

self-incompatible ornamental plants in the Brassicaceae. Sakamoto K, Kusaba M, Nishio T; Mol Gen Genet 1998;258:397-403.

5 603. (sdh cyt) Succinate dehydrogenase cytochrome b subunit signatures

Succinate dehydrogenase (SDH) is a membrane-bound complex of two main components: a membrane-extrinsic component composed of an FAD-binding flavoprotein and an iron-sulfur protein, and a hydrophobic component composed of a cytochrome B and a membrane anchor protein. The cytochrome b component is a mono heme transmembrane protein [1,2,3]

10 belonging to a family that groups: - Cytochrome b-556 from bacterial SDH (gene sdhC). - Cytochrome b560 from the mammalian mitochondrial SDH complex. - Cytochrome b560 subunit encoded in the mitochondrial genome of some algae and in the plant *Marchantia polymorpha*. - Cytochrome b from yeast mitochondrial SDH complex (gene SDH3 or CYB3). - Protein cyt-1 from *Caenorhabditis*. These cytochromes are proteins of about 130 residues
15 that comprise three transmembrane regions. There are two conserved histidines which may be involved in binding the heme group. Two signature patterns have been developed that include these histidine residues.

Consensus pattern: R-P-[LIVMT][LIVMT SEQ ID NO:1])-x(3)-[LIVM][LIVM SEQ ID NO:4])-x(6)-[LIVMWPK][LIVMWPK SEQ ID NO:553])-x(4)-S-x(2)-H-R-x- [ST] [H could
20 be a heme ligand]

Consensus pattern: H-x(3)-[GA]-[LIVMT][LIVMT SEQ ID NO:1])-R-[HF]-[LIVMF][LIVMF SEQ ID NO:2])-x-[FYWM][FYWM SEQ ID NO:137])-D-x-[GVA] [H
could be a heme ligand]

[1] Yu L., Wei Y.-Y., Usui S., Yu C.-A. J. Biol. Chem. 267:24508-24515(1992).[2]

25 Abraham P.R., Mulder A., Van't Riet J., Raue H.A. Mol. Gen. Genet. 242:708-716(1994).[3] Leblanc C., Boyen C., Richard O., Bonnard G., Grienemberger J.M., Kloareg B. J. Mol. Biol. 250:484-495(1995).

30 604. Sec1 family

[1] The Sec1 family: a novel family of proteins involved in synaptic transmission and general secretion. Halachmi N, Lev Z; J Neurochem 1996;66:889-897.

Number of members: 40

605. Protein secE/sec61-gamma signature

In bacteria, the secE protein plays a role in protein export; it is one of the components - with secY and secA - of the preprotein translocase. In eukaryotes, the evolutionary related protein sec61-gamma plays a role in protein translocation through the endoplasmic reticulum; it is part of a trimeric complex that also consists of sec61-alpha and beta [1]. Both secE and sec61-gamma are small proteins of about 60 to 90 amino acids that contain a single transmembrane region at their C-terminal extremity (Escherichia coli secE is an exception, in that it possesses an extra N-terminal segment of 60 residues that contains two additional transmembrane domains). The sequence of secE/sec61-gamma is not extremely well conserved, however it is possible to derive a signature pattern centered on a conserved proline located 10 residues before the beginning of the transmembrane domain.

Consensus pattern: [LIVMFY][LIVMFY SEQ ID NO:18]-x(2)-[DENQGA][DENQGA SEQ ID NO:554]-x(4)-[LIVMFTA][LIVMFTA SEQ ID NO:386]-x-[KRV]-x(2)-[KW]-P-x(3)-[SEQ]-x(7)-[LIVT][LIVT SEQ ID NO:165]-[LIVGA][LIVGA SEQ ID NO:555]-[LIVFGAST][LIVFGAST SEQ ID NO:556]

[1] Hartmann E., Sommer T., Prehn S., Goerlich D., Jentsch S., Rapoport T.A. Nature 367:654-657(1994).

606. 11-S plant seed storage proteins signature

Plant seed storage proteins, whose principal function appears to be the major nitrogen source for the developing plant, can be classified, on the basis of their structure, into different families. 11-S are non-glycosylated proteins which form hexameric structures [1,2]. Each of the subunits in the hexamer is itself composed of an acidic and a basic chain derived from a single precursor and linked by a disulfide bond. This structure is shown in the following representation. +-----+ ||

xxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxNGxCxxxxxxxxxxxxxxxxxxxxxxxx ***** <--

---Acidic-subunit-----><---Basic-subunit-----> <-----About-480-to-500-

residues----->'C': conserved cysteine involved in a disulfide bond.'*': position of the pattern. Proteins that belong to the 11-S family are: pea and broad bean legumins, rape cruciferin, rice glutelins, cotton beta-globulins, soybean glycinins, pumpkin 11-S globulin,

oat globulin, sunflower helianthinin G3, etc. The region that includes the conserved cleavage site between the acidic and basic subunits (Asn-Gly) and a proximal cysteine residue which is involved in the interchain disulfide bond have been used as a signature pattern for this family of proteins.

Consensus pattern: N-G-x-[DE](2)-x-[LIVMF][LIVMF SEQ ID NO:2]]-C-[ST]-x(11,12)-[PAG]-D [C is involved in a disulfide bond

[1] Hayashi M., Mori H., Nishimura M., Akazawa T., Hara-Nishimura I. Eur. J. Biochem. 172:627-632(1988).[2] Shotwell M.A., Afonso C., Davies E., Chesnut R.S., Larkins B.A. Plant Physiol. 87:698-704(1988).

607. 7S seed storage protein

7S globulin is one of the main storage proteins of most angiosperms and gymnosperms. The 7S storage proteins are homotrimers.

Number of members: 67

[1] The three-dimensional structure of canavalin from jack bean (*Canavalia ensiformis*). Ko TP, Ng JD, McPherson A; Plant Physiol 1993;101:729-744.

608. Aspartate-semialdehyde dehydrogenase signature

Aspartate-semialdehyde dehydrogenase (ASD) catalyzes the second step in the common biosynthetic pathway leading from Asp to diaminopimelate and Lys, to Met, and to Thr; the NADP-dependent reductive dephosphorylation of L-aspartyl phosphate to L-aspartate-semialdehyde. In bacteria and fungi, ASD is a protein of about 40 Kd (340 to 370 residues) whose sequence is not extremely well conserved [1]. A conserved cysteine residue has been implicated as important for the catalytic activity [2]. The region of conservation around the active site residue is too small to be used as signature pattern. Another more conserved region, located in the last third of the sequence, and which contains both a conserved cysteine as well as an histidine has been used instead.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4)]-[SADN][SADN SEQ ID NO:71)]-x(2)-C-x-R-[LIVM][LIVM SEQ ID NO:4)]-x(4)-[GSC]-H-[STA

[1] Baril C., Richaud C., Fourni E., Baranton G., Saint Girons I. J. Gen. Microbiol. 138:47-53(1992).[2] Karsten W.E., Viola R.E. Biochim. Biophys. Acta 1121:234-238(1992).

N-acetyl-gamma-glutamyl-phosphate reductase active site

N-acetyl-gamma-glutamyl-phosphate reductase (EC 1.2.1.38) (AGPR) [1,2] is the enzyme that catalyzes the third step in the biosynthesis of arginine from glutamate, the NADP-dependent reduction of N-acetyl-5-glutamyl phosphate into N-acetylglutamate 5-semialdehyde. In bacteria it is a monofunctional protein of 35 to 38 Kd (gene argC) while in fungi it is part of a bifunctional mitochondrial enzyme (gene ARG5,6, arg11 or arg-6) which contains a N-terminal acetylglutamate kinase (EC 2.7.2.8) domain and a C-terminal AGPR domain. In the Escherichia coli enzyme, a cysteine has been shown to be implicated in the catalytic activity, the region around this residue is well conserved and can be used as a signature pattern.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-[GSA]-x-P-G-C-[FY]-[AVP]-T-[GA]-x(3)-[GTAC][GTAC SEQ ID NO:557]-[LIVM][LIVM SEQ ID NO:4]- x-P [C is the active site residue]

[1] Ludovice M., Martin J.F., Carrachas P., Liras P. J. Bacteriol. 174:4606-4613(1992). [2] Gessert S.F., Kim J.H., Nargang F.E., Weiss R.L. J. Biol. Chem. 269:8189-8203(1994).

609. Sialyltransferase family,

Number of members: 18

610. SpoU rRNA Methylase family

This family of proteins probably use S-AdoMet. Number of members: 58

[1] SpoU protein of Escherichia coli belongs to a new family of putative rRNA methylases. Koonin EV, Rudd KE; Nucleic Acids Res 1993;21:5519-5519. [2] The spoU gene of escherichia coli , the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2' methyltransferase activity. Persson BC, Jager G, Gustafsson C; Nucleic Acids Res 1997;25:4093-4097.

611. Stathmin family signatures

Stathmin [1] (from the Greek 'stathmos' which means relay), is an ubiquitous intracellular protein, present in a variety of phosphorylated forms and which serves as a relay for diverse second messenger pathways. Its expression and phosphorylation are regulated throughout development and in response to extracellular signals regulating cell proliferation, differentiation and function. Stathmin is a highly conserved protein of 149 amino acid residues. Structurally, it consists of an N-terminal domain of about 45 residues followed by a 78 residue alpha-helical domain consisting of a heptad repeat coiled coil structure and a C-terminal domain of 25 residues. Protein SCG10 is a neuron-specific, membrane-associated protein that accumulates in the growth cones of developing neurons. It is highly similar in its sequence to stathmin, but differs in that it contains an additional N-terminal hydrophobic segment of 32 residues which is probably responsible for its interaction with membranes. *Xenopus* protein XB3 is also evolutionary related to stathmin and also contains an additional N-terminal hydrophobic domain [2]. A conserved decapeptide which ends with the first three residues of the coiled coil domain and a second pattern that corresponds to part of the central region of the coiled coil have been selected as signatures for proteins of the stathmin family. Consensus pattern: P-[KRQ]-[KR](2)-[DE]-x-S-L-[EG]-E- Consensus pattern: A-E-K-R-E-H-E-[KR]-E- [1] Sobel A. Trends Biochem. Sci. 16:301-305(1991). [2] Maucuer A., Moreau J., Mechali M., Sobel A. J. Biol. Chem. 268:16420-16429(1993).

612. SUA5/yciO/yrdC family signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast protein SUA5. - *Escherichia coli* hypothetical protein yciO and HI1198, the corresponding *Haemophilus influenzae* protein. - *Escherichia coli* hypothetical protein yrdC and HI0656, the corresponding *Haemophilus influenzae* protein. - *Bacillus subtilis* hypothetical protein ywIC. - *Mycobacterium leprae* hypothetical protein in rfe-hemK intergenic region. - *Methanococcus jannaschii* hypothetical protein MJ0062. These are proteins of from 20 to 46 Kd which contain a number of conserved regions in their N-terminal section. They can be picked up in the database by the following pattern.

Consensus pattern: [LIVMTA][LIVMTA SEQ ID NO:311]](3)-[LIVMFYC][LIVMFYC SEQ ID NO:6]]-[PG]-T-[DE]-[STA]-x-[FY]-[GA]-[LIVM][LIVM SEQ ID NO:4]-[GS]-

[1] Bairoch A., Rudd K.E., Robison K. Unpublished observations (1995).

5 613. Sucrose synthase

Sucrose synthases catalyse the synthesis of sucrose from UDP-glucose and fructose. This family includes the bulk of the sucrose synthase protein. However the carboxyl terminal region of the sucrose synthases belongs to the glycosyl transferase family Glycos_transf_1.

10

614. Sulfotransferase proteins

Number of members: 59

15 615. Synaptophysin / synaptoporin signature

Synaptophysin and synaptoporin [1] are structurally related proteins, found in the membrane of synaptic vesicles, which may function as ionic or solute channels. These two glycoproteins seem to span the membrane four times. Both their N- and C-termini sequences seem to be cytoplasmically located. As a signature pattern for this family of proteins, a highly conserved region located in the beginning of the first intravesicular loop just after the first transmembrane domain has been selected. This region contains a cysteine residue that may be involved in a disulfide bond.

20

Consensus pattern: L-S-V-[DE]-C-x-N-K-T [C may be involved in a disulfide bond

[1] Knaus P., Marqueze-Pouey B., Scherer H., Betz H. Neuron 5:453-462(1990).

25

616. Syndecans signature

Syndecans [1,2] (from the greek syndein; to bind together) are a family of transmembrane heparan sulfate proteoglycans which are implicated in the binding of extracellular matrix components and growth factors. Syndecans bind a variety of molecules via their heparan sulfate chains and can act as receptors or as co-receptors. Structurally, these proteins consist of four separate domains: a) A signal sequence; b) An extracellular domain (ectodomain) of variable length and whose sequence is not evolutionary conserved in the various forms of

30

syndecans. The ectodomain contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains; c) A transmembrane region; d) A highly conserved cytoplasmic domain of about 30 to 35 residues which could interact with cytoskeletal proteins. The proteins known to belong to this family are: - Syndecan 1. - Syndecan 2 or fibroglycan. - Syndecan 3 or neuroglycan or N-syndecan. - Syndecan 4 or amphiglycan or ryudocan. - *Drosophila* syndecan. - *Caenorhabditis elegans* probable syndecan (F57C7.3). The signature pattern that has been developed for syndecans starts with the last residue of the transmembrane region and includes the first 10 residues of the cytoplasmic domain. This region, which contains four basic residues, could act as a stop transfer site.

Consensus pattern: [FY]-R-[IM]-[KR]-K(2)-D-E-G-S-Y

[1] Bernfield M., Kokenyesi R., Kato M., Hinkes M.T., Spring J., Gallo R.L., Lose E.J. Annu. Rev. Cell Biol. 8:365-393(1992).[2] David G. FASEB J. 7:1023-1030(1993).

617. Syntaxin / epimorphin family signature

The following proteins have been shown to be evolutionary related [1,2,3]: - Epimorphin (or syntaxin 2), a mammalian mesenchymal protein which plays an essential role in epithelial morphogenesis. - Syntaxin 1A (also known as antigen HPC-1) and syntaxin 1B which are synaptic proteins which may be involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 3. - Syntaxin 4, which is potentially involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 5, which mediates endoplasmic reticulum to golgi transport. - Syntaxin 6, which is involved in intracellular vesicle trafficking. - Syntaxin 7. - Yeast PEP12 (or VPS6) which is required for the transport of proteases to the vacuole. - Yeast SED5 which is required for the fusion of transport vesicles with the Golgi complex. - Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma membrane. - Yeast VAM3, which is required for vacuolar assembly. - *Arabidopsis thaliana* protein KNOLLE which may be involved in cytokinesis. - *Caenorhabditis elegans* hypothetical proteins F35C8.4, F48F7.2, F55A11.2 and T01B11.3. The above proteins share the following characteristics: a size ranging from 30 Kd to 40 Kd; a C-terminal extremity which is highly hydrophobic and is probably involved in anchoring the protein to the membrane; a central, well conserved region, which seems to be in a coiled-coil conformation. The pattern specific for this family is based on the most conserved region of the coiled coil domain.

521

Consensus pattern: [RQ]-x(3)-[LIVMA][LIVMA SEQ ID NO:30)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]-[ESH]-x(2)-[LIVMT][LIVMT SEQ ID NO:1)]-x-[DEVMT][DEVMT SEQ ID NO:263)]-[LIVM][LIVM SEQ ID NO:4)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]-[FS]-x(2)-[LIVM][LIVM SEQ ID NO:4)]-x(3)-[LIVT][LIVT SEQ ID NO:165)]-x(2)-Q-[GADEQ][GADEQ SEQ ID NO:558)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]-[DNQT][DNQT SEQ ID NO:559)]-x-[LIVMF][LIVMF SEQ ID NO:2)]-[DESV][DESV SEQ ID NO:560)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]

[1] Bennett M.K., Garcia-Arraras J.E., Elferink L.A., Peterson K., Fleming A.M., Hazuka C.D., Scheller R.H. Cell 74:863-873(1993).[2] Spring J., Kato M., Bernfield M. Trends Biochem. Sci. 18:124-125(1993).[3] Pelham H.R.B. Cell 73:425-426(1993).

618. Sm protein

The U1, U2, U4/U6, and U5 small nuclear ribonucleoprotein particles (snRNPs) involved in pre-mRNA splicing contain seven Sm proteins (B/B', D1, D2, D3, E, F and G) in common, which assemble around the Sm site present in four of the major spliceosomal small nuclear RNAs. These proteins contain a common sequence motif in two segments, Sm1 and Sm2, separated by a short variable linker.

[1] Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R EMBO J 1995;14:2076-2088. [2] Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K; Cell 1999;96:375-387.

619. Skp1 family

[1] Stebbins CE, Kaelin WG Jr, Pavletich NP; Science 1999;284:455-461.

620. Protein secY signatures

The eubacterial secY protein [1] plays an important role in protein export. It interacts with the signal sequences of secretory proteins as well as with two other components of the protein translocation system: secA and secE. SecY is an integral plasma membrane protein of 419 to 492 amino acid residues that apparently contains ten transmembrane segments. Such a structure probably confers to secY a 'translocator' function, providing a channel for periplasmic and outer-membrane precursor proteins. Homologs of secY are found in archaeobacteria [2]. SecY is also encoded in the chloroplast genome of some algae [3] where it could be involved in a prokaryotic-like protein export system across the two membranes of the chloroplast endoplasmic reticulum (CER) which is present in chromophyte and cryptophyte algae. Two signature patterns have been developed for secY proteins. The first corresponds to the second transmembrane region, which is the most conserved section of these proteins. The second spans the C-terminal part of the fourth transmembrane region, a short intracellular loop, and the N-terminal part of the fifth transmembrane region.

Consensus pattern: [GST]-[LIVMF][LIVMF SEQ ID NO:2]](2)-x-[LIVM][LIVM SEQ ID NO:4]]-G-[LIVM][LIVM SEQ ID NO:4]]-x-P-[LIVMFY][LIVMFY SEQ ID NO:18]](2)-x-[AS]-[GSTQ][GSTQ SEQ ID NO:561]]-[LIVMFAT][LIVMFAT SEQ ID NO:562]](3)-Q-[LIVMFA][LIVMFA SEQ ID NO:51]](2)

Consensus pattern: [LIVMFYW][LIVMFYW SEQ ID NO:26]](2)-x-[DE]-x-[LIVMF][LIVMF SEQ ID NO:2]]-[STN]-x(2)-G-[LIVMF][LIVMF SEQ ID NO:2]]-[GST]-[NST]-G-x-[GST]-[LIVMF][LIVMF SEQ ID NO:2]](3)

[1] Ito K. Mol. Microbiol. 6:2423-2428(1992).[2] Auer J., Spicker G., Boeck A. Biochimie 73:683-688(1991).[3] Douglas S.E. FEBS Lett. 298:93-96(1992).

621. (Seed protein) Small hydrophilic plant seed proteins signature. The following small hydrophilic plant seed proteins are structurally related: - Arabidopsis thaliana proteins GEA1 and GEA6. - Cotton late embryogenesis abundant (LEA) protein D-19. - Carrot EMB-1 protein. - Barley LEA proteins B19.1A, B19.1B, B19.3 and B19.4. - Maize late embryogenesis abundant protein Emb564. - Radish late seed maturation protein p8B6. - Rice embryonic abundant protein Emp1. - Sunflower 10 Kd late embryogenesis abundant protein (DS10). - Wheat Em proteins. These proteins contains from 83 to 153 amino acid residues and may play a role[1,2] in equipping the seed for survival, maintaining a minimal level of hydration in the dry organism and preventing the denaturation of cytoplasmic components.

They may also play a role during imbibition by controlling water uptake. As a signature pattern, the best conserved region in the sequence of these proteins has been developed, it is a glycine-rich nonapeptide located in the N-terminal section.-

5 Consensus pattern: G-[EQ]-T-V-V-P-G-G-T-

[1] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. Plant Mol. Biol. 12:475-486(1989).[2] Gaubier P., Raynal M., Hull G., Huestis G.M., Grellet F., Arenas C., Pages M., Delseny M. Mol. Gen. Genet. 238:409-418(1993).

10

622. Serine carboxypeptidases, active sites

All known carboxypeptidases are either metallo carboxypeptidases or serinecarboxypeptidases. The catalytic activity of the serine carboxypeptidases, like that of the trypsin family serine proteases, is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which is itself hydrogen-bonded to a serine [1]. Proteins known to be serine carboxypeptidases are: - Barley and wheat serine carboxypeptidases I, II, and III [2]. - Yeast carboxypeptidase Y (YSCY) (gene PRC1), a vacuolar protease involved in degrading small peptides. - Yeast KEX1 protease, involved in killer toxin and alpha-factor precursor processing. - Fission yeast *sxa2*, a probable carboxypeptidase involved in degrading or processing mating pheromones [3]. - *Penicillium janthinellum* carboxypeptidase S1 [4]. - *Aspergillus niger* carboxypeptidase pepF. - *Aspergillus sato*i carboxypeptidase cpdS. - Vertebrate protective protein / cathepsin A [5], a lysosomal protein which is not only a carboxypeptidase but also essential for the activity of both beta-galactosidase and neuraminidase. - Mosquito vitellogenic carboxypeptidase (VCP) [6]. - *Naegleria fowleri* virulence-related protein Nf314 [7]. - Yeast hypothetical protein YBR139w. - *Caenorhabditis elegans* hypothetical proteins C08H9.1, F13D12.6, F32A5.3, F41C3.5 and K10B2.2. This family also includes: - Sorghum (s)-hydroxymandelonitrile lyase (hydroxynitrile lyase) (HNL) [8], an enzyme involved in plant cyanogenesis. The sequences surrounding the active site serine and histidine residues are highly conserved in all these serine carboxypeptidases.

30

Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-x-[GTA]-E-S-Y-[AG]-[GS] [S is the active site residue]

Consensus pattern: [LIVF][LIVE SEQ ID NO:127)]-x(2)-[LIVSTA][LIVSTA SEQ ID NO:563)]-x-[IVPST][IVPST SEQ ID NO:564)]-x-[GSDNQL][GSDNQL SEQ ID NO:565)]-[SAGV][SAGV SEQ ID NO:25)]-[SG]-H-x-[IVAQ][IVAQ SEQ ID NO:566)]-P-x(3)-[PSA]
 [H is the active site residue]

- 5 [1] Liao D.I., Remington S.J. J. Biol. Chem. 265:6528-6531(1990).[2] Sorensen S.B., Svendsen I., Breddam K. Carlsberg Res. Commun. 54:193-202(1989).[3] Imai Y., Yamamoto M. Mol. Cell. Biol. 12:1827-1834(1992).[4] Svendsen I., Hofmann T., Endrizzi J., Remington J., Breddam K. FEBS Lett. 333:39-43(1993).[5] Galjart N.J., Morreau H., Willemsen R., Gillemans N., Bonten E.J., d'Azzo A. J. Biol. Chem. 266:14754-14762(1991).[10 6] Cho W.L., Deitsch K.W., Raikhel A.S. Proc. Natl. Acad. Sci. U.S.A. 88:10821-10824(1991).[7] Hu W.N., Kopachik W., Band R.N. Infect. Immun. 60:2418-2424(1992).[8] Wajant H., Mundry K.W., Pfitzenmaier K. Plant Mol. Biol. 26:735-746(1994).[9] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

15

623. Serpins signature. Serpins (SERine Proteinase INhibitors) [1,2,3,4] are a group of structurally related proteins. They are high molecular weight (400 to 500 amino acids),extracellular, irreversible serine protease inhibitors with a well defined structural-functional characteristic: a reactive region that acts as a 'bait' for an appropriate serine protease. This region is found in the C-terminal part of these proteins. Proteins which are known to belong to the serpin family are listed below (references are only provided for recently determined sequences): - Alpha-1 protease inhibitor (alpha-1-antitrypsin, contrapsin). - Alpha-1-antichymotrypsin, - Antithrombin III. - Alpha-2-antiplasmin. - Heparin cofactor II. - Complement C1 inhibitor. - Plasminogen activator inhibitors 1 (PAI-1) and 2 (PAI-2). - Glia derived nexin (GDN) (Protease nexin I). - Protein C inhibitor. - Rat hepatocytes SPI-1, SPI-2 and SPI-3 inhibitors. - Human squamous cell carcinoma antigen (SCCA) which may act in the modulation of the host immune response against tumor cells. - A lepidopteran protease inhibitor. - Leukocyte elastase inhibitor which, in contrast to other serpins, is an intracellular protein. - Neuroserpin [5], a neuronal inhibitor of plasminogen activators and plasmin. - Cowpox virus crmA [6], an inhibitor of the thiol protease interleukin-1B converting enzyme (ICE). CrmA is the only serpin known to inhibit a non-serine proteinase. - Some orthopoxviruses probable protease inhibitors, which may be involved in the regulation of the blood clotting cascade and/or of the complement cascade in
- 20
- 25
- 30

the mammalian host. On the basis of strong sequence similarities, a number of proteins with no known inhibitory activity are said to belong to this family: - Birds ovalbumin and the related genes X and Y proteins. - Angiotensinogen; the precursor of the angiotensin active peptide. - Barley protein Z; the major endosperm albumin. - Corticosteroid binding globulin (CBG). - Thyroxine-binding globulin (TBG). - Sheep uterine milk protein (UTMP) and pig uteroferrin-associated protein (UFAP). - Hsp47, an endoplasmic reticulum heat-shock protein that binds strongly to collagen and could act as a chaperone in the collagen biosynthetic pathway [7]. - Maspin, which seems to function as a tumor suppressor [5]. - Pigment epithelium-derived factor precursor (PEDF), a protein with a strong neutrophilic activity [8]. - Ep45, an estrogen-regulated protein from *Xenopus* [9]. A signature pattern has been developed for this family of proteins, centered on a well conserved Pro-Phe sequence which is found ten to fifteen residues on the C-terminal side of the reactive bond

Consensus pattern: [LIVMFY][LIVMFY SEQ ID NO:18]]-x-[LIVMFYAC][LIVMFYAC SEQ ID NO:97]]-[DNQ]-[RKHQSS][RKHQSS SEQ ID NO:567]]-[PST]-F-
[LIVMFY][LIVMFY SEQ ID NO:18]]-[LIVMFYC][LIVMFYC SEQ ID NO:6]]-x-
[LIVMFAH][LIVMFAH SEQ ID NO:568]]-

[1] Carrell R., Travis J. Trends Biochem. Sci. 10:20-24(1985).[2] Carrell R., Pemberton P.A., Boswell D.R. Cold Spring Harbor Symp. Quant. Biol. 52:527-535(1987).[3] Huber R., Carrell R.W. Biochemistry 28:8951-8966(1989).[4] Remold-O'Donneel E. FEBS Lett. 315:105-108(1993).[5] Osterwalder T., Contartese J., Stoeckli E.T., Kuhn T.B., Sonderegger P. EMBO J. 15:2944-2953(1996).[6] Komiyama T., Ray C.A., Pickup D.J., Howard A.D., Thornberry N.A., Peterson E.P., Salvesen G. J. Biol. Chem. 269:19331-19337(1994).[7] Clarke E., Sandwal B.D. Biochim. Biophys. Acta 1129:246-248(1992).[8] Zou Z., Anisowicz A., Neveu M., Rafidi K., Sheng S., Sager R., Hendrix M.J., Seftor E., Thor A. Science 263:526-529(1994).[9] Steele F.R., Chader G.J., Johnson L.V., Tombran-Tink J. Proc. Natl. Acad. Sci. U.S.A. 90:1526-1530(1993).[10] Holland L.J., Suksang C., Wall A.A., Roberts L.R., Moser D.R., Bhattacharya A. J. Biol. Chem. 267:7053-7059(1992).

Some bacterial regulatory proteins activate the expression of genes from promoters recognized by core RNA polymerase associated with the alternative sigma-54 factor. These have a conserved domain of about 230 residues involved in the ATP-dependent [1,2] interaction with sigma-54. This domain has been found in the proteins listed below:

- *acoR* from *Alcaligenes eutrophus*, an activator of the acetoin catabolism operon *acoXABC*.
- *algB* from *Pseudomonas aeruginosa*, an activator of alginate biosynthetic gene *algD*.
- *dctD* from *Rhizobium*, an activator of *dctA*, the C4-dicarboxylate transport protein.
- *dhaR* from *Citrobacter freundii*, a regulator of the *dha* operon for glycerol utilization.
- *fhlA* from *Escherichia coli*, an activator of the formate dehydrogenase H and hydrogenase III structural genes.
- *flbD* from *Caulobacter crescentus*, an activator of flagellar genes.
- *hoxA* from *Alcaligenes eutrophus*, an activator of the hydrogenase operon.
- *hrpS* from *Pseudomonas syringae*, an activator of *hprD* as well as other *hrp* loci involved in plant pathogenicity.
- *hupR1* from *Rhodobacter capsulatus*, an activator of the [NiFe] hydrogenase genes *hupSL*.
- *hydG* from *Escherichia coli* and *Salmonella typhimurium*, an activator of the hydrogenase activity.
- *levR* from *Bacillus subtilis*, which regulates the expression of the levanase operon (*levDEFG* and *sacC*).
- *nifA* (as well as *anfA* and *vnfA*) from various bacteria, an activator of the *nif* nitrogen-fixing operon.
- *ntrC*, from various bacteria, an activator of nitrogen assimilatory genes such as that for glutamine synthetase (*glnA*) or of the *nif* operon.
- *pgtA* from *Salmonella typhimurium*, the activator of the inducible phospho- glycerate transport system.
- *pilR* from *Pseudomonas aeruginosa*, an activator of pilin gene transcription.
- *rocR* from *Bacillus subtilis*, an activator of genes for arginine utilization
- *tyrR* from *Escherichia coli*, involved in the transcriptional regulation of aromatic amino-acid biosynthesis and transport.
- *wtsA*, from *Erwinia stewartii*, an activator of plant pathogenicity gene *wtsB*.
- *xylR* from *Pseudomonas putida*, the activator of the *tol* plasmid xylene catabolism operon *xylCAB* and of *xylS*.
- *Escherichia coli* hypothetical protein *yfhA*.
- *Escherichia coli* hypothetical protein *yhgB*.

About half of these proteins (*algB*, *dcdT*, *flbD*, *hoxA*, *hupR1*, *hydG*, *ntrC*, *pgtA* and *pilR*) belong to signal transduction two-component systems [3] and possess a domain that can be phosphorylated by a sensor-kinase protein in their N- terminal section. Almost all of these proteins possess a helix-turn-helix DNA-binding domain in their C-terminal section. The domain which interacts with the sigma-54 factor has an ATPase activity. This may be required to promote a conformational change necessary for the interaction [4]. The domain contains an atypical ATP-binding motif A (P-loop) as well as a form of motif B. The two ATP-binding motifs are located in the N-terminal section of the

domain; signature patterns have been developed for both motifs. Other regions of the domain are also conserved. One of them, located in the C-terminal section, has been selected as a third signature pattern.

Consensus pattern: [LIVMFY][LIVMFY SEQ ID NO:18]](3)-x-G-[DEQ]-[STE]-G-
5 [STAV][STAV SEQ ID NO:105]]-G-K-x(2)-[LIVMFY][LIVMFY SEQ ID NO:18]]

Consensus pattern: [GS]-x-[LIVMF][LIVMF SEQ ID NO:2]]-x(2)-A-
[DNEQASH][DNEQASH SEQ ID NO:569]]-[GNEK][GNEK SEQ ID NO:570]]-G-
[SFIM][SFIM SEQ ID NO:571]]-[LIVMFY][LIVMFY SEQ ID NO:18]](3)-[DE]-[EK]-
[LIVM][LIVM SEQ ID NO:4]]

10 Consensus pattern: [FYW]-P-[GS]-N-[LIVM][LIVM SEQ ID NO:4]]-R-[EQ]-L-x-
[NHAT][NHAT SEQ ID NO:572]]

[1] Morrett E., Segovia L. J. Bacteriol. 175:6067-6074(1993).[2] Austin S., Kundrot C.,
Dixon R. Nucleic Acids Res. 19:2281-2287(1991).[3] Albright L.M., Huala E., Ausubel
F.M. Annu. Rev. Genet. 23:311-336(1989).[4] Austin S., Dixon R. EMBO J. 11:2219-
15 2228(1992).

625. Sigma-70 factors family signatures

Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of
20 the core RNA polymerase to specific initiation sites and are then released. They alter the
specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors.
Two of these factors, sigma-70 (gene rpoD), generally known as the major or primary sigma
factor, and sigma-54 (gene rpoN or ntrA) direct the transcription of a wide variety of genes.
The other sigma factors, known as alternative sigma factors, are required for the transcription
25 of specific subsets of genes. With regard to sequence similarity, sigma factors can be grouped
into two classes: the sigma-54 and sigma-70 families. The sigma-70 family includes, in
addition to the primary sigma factor, a wide variety of sigma factors, some of which are listed
below: - Bacillus sigma factors involved in the control of sporulation-specific genes: sigma-E
(sigE or spoIIGB), sigma-F (sigF or spoIIAC), sigma-G (sigG or spoIIIG), sigma-H (sigH or
30 spo0C) and sigma-K (sigK or spoIVCB/spoIIIC). - Escherichia coli and related bacteria
sigma-32 (gene rpoH or htpR) involved in the expression of heat shock genes. - Escherichia
coli and related bacteria sigma-27 (gene fliA) involved in the expression of the flagellin gene.
- Escherichia coli sigma-S (gene rpoS or katF) which seems to be involved in the expression

of genes required for protection against external stresses. - *Myxococcus xanthus* sigma-B (sigB) which is essential for the late-stage differentiation of that bacteria. Alignments of the sigma-70 family permit the identification of four regions of high conservation [2,3]. Each of these four regions can in turn be subdivided into a number of sub-regions. Signature patterns based on the two best-conserved sub-regions have been developed. The first pattern corresponds to sub-region 2.2; the exact function of this sub-region is not known although it could be involved in the binding of the sigma factor to the core RNA polymerase. The second pattern corresponds to sub-region 4.2 which seems to harbor a DNA-binding 'helix-turn-helix' motif involved in binding the conserved -35 region of promoters recognized by the major sigma factors. The second pattern starts one residue before the N-terminal extremity of the HTH region and ends six residues after its C-terminal extremity.

Consensus pattern: [DE]-[LIVMF][LIVMF SEQ ID NO:2](2)-[HEQS][HEQS SEQ ID NO:573]-x-G-x-[LIVMFA][LIVMFA SEQ ID NO:81]-G-L-[LIVMFYE][LIVMFYE SEQ ID NO:574]-x-[GSAM][GSAM SEQ ID NO:575]-[LIVMAP][LIVMAP SEQ ID NO:253]

Consensus pattern: [STN]-x(2)-[DEQ]-[LIVM][LIVM SEQ ID NO:4]-[GAS]-x(4)-[LIVMF][LIVMF SEQ ID NO:2]-[PSTG][PSTG SEQ ID NO:576]-x(3)-[LIVMA][LIVMA SEQ ID NO:30]-x-[NQR]-[LIVMA][LIVMA SEQ ID NO:30]-[EQH]-x(3)-[LIVMFW][LIVMFW SEQ ID NO:13]-x(2)-[LIVM][LIVM SEQ ID NO:4]

[1] Helmann J.D., Chamberlin M.J. Annu. Rev. Biochem. 57:839-872(1988).[2] Gribskov M., Burgess R.R. Nucleic Acids Res. 14:6745-6763(1986).[3] Lonetto M.A., Gribskov M., Gross C.A. J. Bacteriol. 174:3843-3849(1992).[4] Lonetto M.A., Brown K.L., Rudd K.E., Buttner M.J. Proc. Natl. Acad. Sci. U.S.A. 91:7573-7577(1994).

626. Signal carboxyl-terminal domain. 430 members.

627. Signal peptidases I signatures

Signal peptidases (SPases) [1] (also known as leader peptidases) remove the signal peptides from secretory proteins. In prokaryotes three types of Spases are known: type I (gene *lepB*) which is responsible for the processing of the majority of exported pre-proteins; type II (gene *lsp*) which only process lipoproteins, and a third type involved in the processing of pili subunits. SPase I is an integral membrane protein that is anchored in the cytoplasmic

membrane by one (in *B. subtilis*) or two (in *E. coli*) N-terminal transmembrane domains with the main part of the protein protruding in the periplasmic space. Two residues have been shown [2,3] to be essential for the catalytic activity of SPase I: a serine and an lysine. SPase I is evolutionary related to the yeast mitochondrial inner membrane protease subunit 1 and 2 (genes IMP1 and IMP2) which catalyze the removal of signal peptides required for the targeting of proteins from the mitochondrial matrix, across the inner membrane, into the inter-membrane space [4]. In eukaryotes the removal of signal peptides is effected by an oligomeric enzymatic complex composed of at least five subunits: the signal peptidase complex (SPC). The SPC is located in the endoplasmic reticulum membrane. Two components of mammalian SPC, the 18 Kd (SPC18) and the 21 Kd (SPC21) subunits as well as the yeast SEC11 subunit have been shown [5] to share regions of sequence similarity with prokaryotic SPases I and yeast IMP1/IMP2. Three signature patterns for these proteins have been developed. The first signature contains the putative active site serine, the second signature contains the putative active site lysine which is not conserved in the SPC subunits, and the third signature corresponds to a conserved region of unknown biological significance which is located in the C-terminal section of all these proteins.

Consensus pattern: [GS]-x-S-M-x-[PS]-[AT]-[LF] [S is an active site residue]

Consensus pattern: K-R-[LIVMSTA][LIVMSTA SEQ ID NO:433](2)-G-x-[PG]-G-[DE]-x-[LIVM][LIVM SEQ ID NO:4]-x-[LIVMFY][LIVMFY SEQ ID NO:18] [K is an active site residue]

Consensus pattern: [LIVMFYW][LIVMFYW SEQ ID NO:26](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG]

[1] Dalbey R.E., von Heijne G. Trends Biochem. Sci. 17:474-478(1992).[2] Sung M., Dalbey R.E. J. Biol. Chem. 267:13154-13159(1992).[3] Black M.T. J. Bacteriol. 175:4957-4961(1993).[4] Nunnari J., Fox T.D., Walter P. Science 262:1997-2004(1993).[5] van Dijk J.M., de Jong A., Vehmaanpera J., Venema G., Bron S. EMBO J. 11:2819-2828(1992).[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

628. (sodcu) Copper/Zinc superoxide dismutase signatures

Copper/Zinc superoxide dismutase (SODC) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. SODC binds one atom each of zinc and copper. Various forms of SODC are known: acytoplasmic form in eukaryotes, an additional

chloroplast form in plants, an extracellular form in some eukaryotes, and a periplasmic form in prokaryotes. The metal binding sites are conserved in all the known SODC sequences [2]. Two signature patterns have been derived for this family of enzymes: the first one contains two histidine residues that bind the copper atom; the second one is located in the C-terminal section of SODC and contains a cysteine which is involved in a disulfide bond.

Consensus pattern: [GA]-[~~IMFAT~~][IMFAT SEQ ID NO:577]-H-[~~LIVE~~][LIVE SEQ ID NO:127]-H-x(2)-[GP]-[SDG]-x-[~~STAGDE~~][STAGDE SEQ ID NO:578] [The two H's are copper ligands]

Consensus pattern: G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV] [C is involved in a disulfide bond]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).
[2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

629. (sodfe) Manganese and iron superoxide dismutases signature

Manganese superoxide dismutase (SODM) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. The four ligands of the manganese atom are conserved in all the known SODM sequences. These metal ligands are also conserved in the related iron form of superoxide dismutases [2,3]. A short conserved region which includes two of the four ligands: an aspartate and a histidine has been selected as a signature.

Consensus pattern: D-x-W-E-H-[STA]-[FY](2) [D and H are manganese/iron ligands]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).
[2] Parker M.W., Blake C.C.F. FEBS Lett. 229:377-382(1988). [3] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

630. Spectrin repeat

Spectrin repeats are found in several proteins involved in cytoskeletal structure. These include spectrin, alpha-actinin and dystrophin. The sequence repeat used in this family is taken from the structural repeat in reference [2]. The spectrin repeat forms a three helix bundle. The second helix is interrupted by proline in some sequences.

Number of members: 898

[1] Actin-binding proteins. 1: Spectrin super family. Hartwig JH; Protein Profile 1995;2:732-732. [2] Crystal structure of the repetitive segments of spectrin. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, Branton D; Science 1993;262:2027-2030.

5

631. (subtilase) Streptomyces subtilisin-type inhibitors signature

Bacteria of the Streptomyces family produce a family of proteinase inhibitors[1] characterized by their strong activity toward subtilisin. They are collectively known as SSI's: Streptomyces Subtilisin Inhibitors. Some SSI's also inhibit trypsin or chymotrypsin. In their mature secreted form, SSI's are proteins of about 110 residues with two conserved disulfide bonds. +-----+ +-----+ |||

10

xxxxxxxxxxxxxxxxCxxxxxxxxCxxxxxxxxCx#xxxxxxxxxxxxxxxxCxxxxxx *****'C': conserved cysteine involved in a disulfide bond. '#': active site residue. '*': position of the pattern.

15

Consensus pattern: C-x-P-x(2,3)-G-x-H-P-x(4)-A-C-[ATD]-x-L [The two C's are involved in a disulfide bond]

[1] Taguchi S., Kojima S., Terabe M., Miura K.-I., Momose H. Eur. J. Biochem. 220:911-918(1994).

20

632. Sugar transport proteins signatures

In mammalian cells the uptake of glucose is mediated by a family of closely related transport proteins which are called the glucose transporters [1,2,3]. At least seven of these transporters are currently known to exist (in Human they are encoded by the GLUT1 to GLUT7

25

genes). These integral membrane proteins are predicted to comprise twelve membrane spanning domains. The glucose transporters show sequence similarities [4,5] with a number of other sugar or metabolite transport proteins listed below (references are only provided for recently determined sequences). - Escherichia coli arabinose-proton symport (araE). -

30

Escherichia coli galactose-proton symport (galP). - Escherichia coli and Klebsiella pneumoniae citrate-proton symport (also known as citrate utilization determinant) (gene cit). - Escherichia coli alpha-ketoglutarate permease (gene kgtP). - Escherichia coli proline/betaine transporter (gene proP) [6]. - Escherichia coli xylose-proton symport (xyle). - Zymomonas mobilis glucose facilitated diffusion protein (gene glf). - Yeast high and low

- affinity glucose transport proteins (genes SNF3, HXT1 to HXT14). - Yeast galactose transporter (gene GAL2). - Yeast maltose permeases (genes MAL3T and MAL6T). - Yeast myo-inositol transporters (genes ITR1 and ITR2). - Yeast carboxylic acid transporter protein homolog JEN1. - Yeast inorganic phosphate transporter (gene PHO84). - *Kluyveromyces* lactis lactose permease (gene LAC12). - *Neurospora crassa* quinate transporter (gene Qa-y), and *Emericella nidulans* quinate permease (gene qutD). - *Chlorella* hexose carrier (gene HUP1). - *Arabidopsis thaliana* glucose transporter (gene STP1). - Spinach sucrose transporter. - *Leishmania donovani* transporters D1 and D2. - *Leishmania enriettii* probable transport protein (LTP). - Yeast hypothetical proteins YBR241c, YCR98c and YFL040w. - *Caenorhabditis elegans* hypothetical protein ZK637.1. - *Escherichia coli* hypothetical proteins yabE, ydjE and yhjE. - *Haemophilus influenzae* hypothetical proteins HI0281 and HI0418. - *Bacillus subtilis* hypothetical proteins yxbC and yxdF. It has been suggested [4] that these transport proteins have evolved from the duplication of an ancestral protein with six transmembrane regions, this hypothesis is based on the conservation of two G-R-[KR] motifs. The first one is located between the second and third transmembrane domains and the second one between transmembrane domains 8 and 9. Two patterns have been developed to detect this family of proteins. The first pattern is based on the G-R-[KR] motif; but because this motif is too short to be specific to this family of proteins, a pattern from a larger region centered on the second copy of this motif was derived. The second pattern is based on a number of conserved residues which are located at the end of the fourth transmembrane segment and in the short loop region between the fourth and fifth segments.
- Consensus pattern: [LIVMSTAG][LIVMSTAG SEQ ID NO:44]]-
[LIVMFSAG][LIVMFSAG SEQ ID NO:579]]-x(2)-[LIVMSA][LIVMSA SEQ ID
NO:187]]-[DE]-x-[LIVMFYWA][LIVMFYWA SEQ ID NO:41]]-G- R-[RK]-x(4,6)-
[GSTA][GSTA SEQ ID NO:19]]
- Consensus pattern: [LIVMF][LIVMF SEQ ID NO:2]]-x-G-[LIVMFA][LIVMFA SEQ ID
NO:81]]-x(2)-G-x(8)-[LIFY][LIFY SEQ ID NO:580]]-x(2)-[EQ]-x(6)- [RK]
- [1] Silverman M. Annu. Rev. Biochem. 60:757-794(1991).[2] Gould G.W., Bell G.I. Trends Biochem. Sci. 15:18-23(1990).[3] Baldwin S.A. Biochim. Biophys. Acta 1154:17-49(1993).[4] Maiden M.C.J., Davis E.O., Baldwin S.A., Moore D.C.M., Henderson P.J.F. Nature 325:641-643(1987).[5] Henderson P.J.F. Curr. Opin. Struct. Biol. 1:590-601(1991).[6] Culham D.E., Lasby B., Marangoni A.G., Milner J.L., Steer B.A., van Nues R.W., Wood J.M. J. Mol. Biol. 229:268-276(1993).

633. Synaptobrevin signature

Synaptobrevin [1] is an intrinsic membrane protein of small synaptic vesicles whose function is not yet known, but which is highly conserved in mammals, electric ray (where its is known as VAMP-1), Drosophila and yeast [2]. In yeast there are two closely related forms of synaptobrevin (genes SNC1 and SNC2) while in mammals there is at least 4 (genes SYB1, SYB2, SYB3 and SYBL1). Structurally synaptobrevin consist of a N-terminal cytoplasmic domain of from 90 to 110 residues, followed by a transmembrane region, and then by a short (from 2 to 22 residues) C-terminal intravesicular domain. As a signature pattern for synaptobrevin, a highly conserved stretch of residues located in the central part of the sequence was selected.

Consensus pattern: N-[LIVM][LIVM SEQ ID NO:4)]-~~[DENS]~~[DENS SEQ ID NO:405)]-[KL]-V-x-[DEQ]-R-x(2)-[KR]-[LIVM][LIVM SEQ ID NO:4)]-~~[STDE]~~[STDE SEQ ID NO:581)]- x-[LIVM][LIVM SEQ ID NO:4)]-x-[DE]-[KR]-[TA]-[DE]

[1] Suedhof T.C., Baumert M., Perin M.S., Jahn R. Neuron 2:1475-1481(1989).[2] Gerst J.E., Rodgers L., Riggs M., Wigler M. Proc. Natl. Acad. Sci. U.S.A. 89:4338-4342(1992).

634. TBC domain. Identification of a TBC domain in GYP6_YEAST and GYP7_YEAST, which are GTPase activator proteins of yeast Ypt6 and Ypt7, imply that these domains are GTPase activator proteins of Rab-like small GTPases. Number of members: 55

[1] Medline: 96032578. Molecular cloning of a cDNA with a novel domain present in the tre-2 oncogene and the yeast cell cycle regulators BUB2 and cdc16. Richardson PM, Zon LI; Oncogene 1995;11:1139-1148.

[2] Medline: 97398935. A shared domain between a spindle assembly checkpoint protein and Ypt/Rab-specific GTPase-activators. Neuwald AF; Trends Biochem Sci 1997;22:243-244.

635. Transcription factor TFIID repeat signature (TBP)

Transcription factor TFIID (or TATA-binding protein, TBP) [1,2] is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II.

TFIID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources. This region is necessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region. The intramolecular symmetry generates a saddle-shaped structure that sits astride the DNA [3]. Drosophila TRF (TBP-related factor) [4] is a sequence-specific transcription factor that also binds to the TATA box and is highly similar to TFIID. Archaeobacteria also possess a TBP homolog [5]. A signature pattern that spans the last 50 residues of the repeated region has been derived.-

Consensus pattern: Y-x-P-x(2)-[IF]-x(2)-[LIVM][LIVM SEQ ID NO:4](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)-L-[LIVM][LIVM SEQ ID NO:4]-F-x-[STN]-G-[KR]-[LIVM][LIVM SEQ ID NO:4]-x(3)-G-[TAGL][TAGL SEQ ID NO:582]-[KR]-x(7)-[AGC]-x(7)-[LIVM

[1] Hoffmann A., Sinn E., Yamamoto T., Wang J., Roy A., Horikoshi M., Roeder R.G.

Nature 346:387-390(1990).[2] Gash A., Hoffmann A., Horikoshi M., Roeder R.G., Chua N.-

H. Nature 346:390-394(1990).[3] Nikolov D.B., Hu S.-H., Lin J., Gasch A., Hoffmann A., Horikoshi M., Chua N.-H., Roeder R.G., Burley S.K. Nature 360:40-46(1992).[4] Crowley

T.E., Hoey T., Liu J.-K., Jan Y.N., Jan L.Y., Tjian R. Nature 361:557-561(1993).[5] Marsh

T.L., Reich C.I., Whitelock R.B., Olsen G.J. Proc. Natl. Acad. Sci. U.S.A. 91:4180-4184(1994).

5 636. Translationally controlled tumor protein signatures (TCTP)

Mammalian translationally controlled tumor protein (TCTP) (or P23) is a protein which has been found to be preferentially synthesized in cells during the early growth phase of some types of tumor [1,2], but which is also expressed in normal cells. The physiological function of TCTP is still not known. It is a hydrophilic protein of 18 to 20 Kd. Close homologs have
10 been found in plants [3], earthworm [4], *Caenorhabditis elegans* (F52H2.11), *Hydra*, budding yeast (YKL056c) [5] and fission yeast (SpAC1F12.02c) Two of the best conserved regions have been selected as signature patterns for TCTP.

Consensus pattern: [IFA]-[GA]-[GAS]-N-[PAK]-S-[GA]-E-[GDE]-[PAGE][PAGE SEQ ID NO:583]-[DEQGA][DEQGA SEQ ID NO:584]

15 Consensus pattern: [FLVH][FLVH SEQ ID NO:585]-[FY]-[IVCT][IVCT SEQ ID NO:586]-G-E-x-[MA]-x(2,5)-[DEN]-[GAST][GAST SEQ ID NO:179]-x-[LV]-[AV]-x(3)-[FYW]

[1] Boehm H., Beendorf R., Gaestel M., Gross B., Nuernberg P., Kraft R., Otto A., Bielka H. Biochem. Int. 19:277-286(1989).[2] Makrides S., Chitpatima S.T., Bandyopadhyay R.,
20 Brawerman G. Nucleic Acids Res. 16:2350-2350(1988).[3] Pay A., Heberle-Bors E., Hirt H. Plant Mol. Biol. 19:501-503(1992).[4] Stuerzenbaum S.R., Kille P., Morgan A.J. Biochim. Biophys. Acta 1398:294-304(1998).[5] Rasmussen S.W. Yeast 10:S63-S68(1994).

25 637. TFIIS zinc ribbon domain signature

Transcription factor S-II (TFIIS) [1] is a eukaryotic protein necessary for efficient RNA polymerase II transcription elongation, past template-encoded pause sites. TFIIS shows DNA-binding activity only in the presence of RNA polymerase II. It is a protein of about 300 amino acids whose sequence is highly conserved in mammals, *Drosophila*, yeast (where it
30 was first known as PPR2, a transcriptional regulator of URA4, and then as DST1, the DNA strand transfer protein alpha [2]) and in the archaebacteria *Sulfolobus acidocaldarius* [3]. This family also includes the eukaryotic and archaebacterial RNA polymerase subunits of the 15 Kd / M family (see <PDOC00790>) as well as the following viral proteins: - Vaccinia virus

RNA polymerase 30 Kd subunit (rpo30) [4]. - African swine fever virus protein I243L [5]. The best conserved region of all these proteins contains four cysteines that bind a zinc ion and fold in a conformation termed a 'zinc ribbon' [6]. Besides these cysteines, there are a number of other conserved residues which can be used to help define a specific pattern for this type of domain.

Consensus pattern: C-x(2)-C-x(9)-[LIVMQSAR][LIVMQSAR SEQ ID NO:587]-[QH]-[STQL][STQL SEQ ID NO:588]-[RA]-[SACR][SACR SEQ ID NO:589]-x-[DE]-[DET]-[PGSEA][PGSEA SEQ ID NO:590]-x(6)-C-x(2,5)-C-x(3)-[FW] [The four C's are zinc ligands]

[1] Hirashima S., Hirai H., Nakanishi Y., Natori S. J. Biol. Chem. 263:3858-3863(1988).[2] Kipling D., Kearsey S.E. Nature 353:509-509(1991).[3] Langer D., Zillig W. Nucleic Acids Res. 21:2251-2251(1993).[4] Ahn B.-Y., Gershon P.D., Jones E.V., Moss B. Mol. Cell. Biol. 10:5433-5441(1990).[5] Rodriguez J.M., Salas M.L., Vinuela E. Virology 186:40-52(1992).[6] Qian X., Jeon C., Yoon H., Agarwal K., Weiss M.A. Nature 365:277-279(1993).

638. Tetrahydrofolate dehydrogenase/cyclohydrolase signatures (THF DHG CYH)

Enzymes that participate in the transfer of one-carbon units are involved in various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by formyltetrahydrofolate synthetase (EC 6.3.4.3), methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5 or EC 1.5.1.15) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9). The dehydrogenase and cyclohydrolase activities are expressed by a variety of multifunctional enzymes: - Eukaryotic C-1-tetrahydrofolate synthase (C1-THF synthase), which catalyzes all three reactions described above. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the dehydrogenase/cyclohydrolase domain is located in the N-terminal section of the 900 amino acids protein and consists of about 300 amino acid residues. The C1-THF synthases are NADP- dependent. - Eukaryotic mitochondrial bifunctional dehydrogenase/cyclohydrolase [2]. This is an homodimeric NAD-dependent enzyme of about 300 amino acid residues. - Bacterial fold [3]. Fold is an homodimeric bifunctional NADP-dependent enzyme of about 290 amino acid residues. The sequence of the dehydrogenase/cyclohydrolase domain is

highly conserved in all forms of the enzyme. Two conserved regions have been selected as signature patterns. The first one is located in the N-terminal part of these enzymes and contains three acidic residues. The second pattern is a highly conserved sequence of 9 amino acids which is located in the C-terminal section.

5 Consensus pattern: [EQ]-x-[EQK]-[~~LIVM~~][LIVM SEQ ID NO:4](2)-x(2)-[~~LIVM~~][LIVM SEQ ID NO:4]-x(2)-[~~LIVMY~~][LIVMY SEQ ID NO:141]-N-x-[DN]- x(5)-[~~LIVMF~~][LIVMF SEQ ID NO:2](3)-Q-L-P-[LV]

Consensus pattern: P-G-G-V-G-P-[MF]-T-[IV]

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).[2] Belanger C.,
10 Mackenzie R.E. J. Biol. Chem. 264:4837-4843(1989).[3] d'Ari L., Rabinowitz J.C. J. Biol. Chem. 266:23953-23958(1991).

639. Triosephosphate isomerase active site (TIM)

15 Triosephosphate isomerase (EC 5.3.1.1) (TIM) [1] is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism [2]. The
20 sequence around the active site residue is perfectly conserved in all known TIM's and can be used as a signature pattern for this type of enzyme.

Consensus pattern: [AV]-Y-E-P-[~~LIVM~~][LIVM SEQ ID NO:4]-W-[SA]-I-G-T-[GK] [E is the active site residue]

[1] Lolis E., Alber T., Davenport R.C., Rose D., Hartman F.C., Petsko G.A. Biochemistry
25 29:6609-6618(1990).[2] Knowles J.R. Nature 350:121-124(1991).

640. Thymidine kinase cellular-type signature (TK)

30 Thymidine kinase (TK) (EC 2.7.1.21) is an ubiquitous enzyme that catalyzes the ATP-dependent phosphorylation of thymidine. A comparison of TK sequences has shown [1,2,3] that there are two different families of TK. One family groups together TK from herpes viruses as well as cellular thymidylate kinases, while the second family currently consists of TK from the following sources: - Vertebrates. - Bacterial. - Bacteriophage T4. - Pox viruses.

- African swine fever virus (ASF). - Fish lymphocystis disease virus (FLDV). A conserved region which is located in the C-terminal section of these enzymes has been selected as a signature pattern for this family of TKA.

Consensus pattern: [GA]-x(1,2)-[DE]-x-Y-x-[STAP][STAP SEQ ID NO:135]-x-C-[NKR]-
 5 x-[CH]-[LIVMFYWH][LIVMFYWH SEQ ID NO:591]

[1] Boyle D.B., Coupar B.E.H., Gibbs A.J., Seigman L.J., Both G.W. Virology 156:355-365(1987).[2] Blasco R., Lopez-Otin C., Munoz M., Bockamp E.-O., Simon-Mateo C., Vinuela E. Virology 178:301-304(1990).[3] Robertson G.R., Whalley J.M. Nucleic Acids Res. 16:11303-11317(1988).

10

641. Thymidine kinase from herpesvirus (TK herpes)

[1]

Medline: 96003730

15

Crystal structures of the thymidine kinase from herpes simplex virus type-1 in complex with deoxythymidine and ganciclovir.

Brown DG, Visse R, Sandhu G, Davies A, Rizkallah PJ, Melitz C, Summers WC, Sanderson MR;

20

Nat Struct Biol 1995;2:876-881.

Number of members: 65

642. Nuclear transition protein 2 signatures (TP2)

25

In mammals, the second stage of spermatogenesis is characterized by the conversion of nucleosomal chromatin to the compact, non-nucleosomal and transcriptionally inactive form found in the sperm nucleus. This condensation is associated with a double-protein transition. The first transition corresponds to the replacement of histones by several spermatid-specific proteins, also called transition proteins, which are themselves replaced by protamines during
 30 the second transition. Nuclear transition protein 2 (TP2) is one of those spermatid-specific proteins. TP2 is a basic, zinc-binding protein [1] of 116 to 137 amino-acid residues.

Structurally, TP2 consists of three distinct parts: a conserved serine-rich N-terminal domain of about 25 residues, a variable central domain of 20 to 50 residues which contains cysteine

residues, and a conserved C-terminal domain of about 70 residues rich in lysines and arginines. Two signature patterns for TP2 have been developed: one located in the N-terminal domain, the other in the C-terminal.

Consensus pattern: H-x(3)-H-S-[NS]-S-x-P-Q-S

5 Consensus pattern: K-x-R-K-x(2)-E-G-K-x(2)-K-[KR]-K

[1] Baskaran R., Rao M.R.S. Biochem. Biophys. Res. Commun. 179:1491-1499(1991).

643. Thiamine pyrophosphate enzymes signature (TTP enzymes)

10 A number of enzymes require thiamine pyrophosphate (TPP) (vitamin B1) as a cofactor. It has been shown [1] that some of these enzymes are structurally related. These related TPP enzymes are: - Pyruvate oxidase (POX) (EC 1.2.3.3) Reaction catalyzed: pyruvate + orthophosphate + O(2) + H(2)O = acetyl phosphate + CO(2) + H(2)O(2). - Pyruvate decarboxylase (PDC) (EC 4.1.1.1) Reaction catalyzed: pyruvate = acetaldehyde + CO(2). -
15 Indolepyruvate decarboxylase (EC 4.1.1.74) [2] Reaction catalyzed: indole-3-pyruvate = indole-3-acetaldehyde + CO(2). - Acetolactate synthase (ALS) (EC 4.1.3.18) Reaction catalyzed: 2 pyruvate = acetolactate + CO(2). - Benzoylformate decarboxylase (BFD) (EC 4.1.1.7) [3] Reaction catalyzed: benzoylformate = benzaldehyde + CO(2). A conserved region which is located in their C-terminal section has been selected as a signature pattern for
20 these enzymes.

Consensus pattern: [LIVMF][LIVMF SEQ ID NO:2]-[GSA]-x(5)-P-x(4)-
[LIVMFYW][LIVMFYW SEQ ID NO:26]-x-[LIVMF][LIVMF SEQ ID NO:2]-x-G-D-
[GSA]-[GSAC][GSAC SEQ ID NO:93]

[1] Green J.B.A. FEBS Lett. 246:1-5(1989).[2] Koga J., Adachi T., Hidaka H. Mol. Gen. Genet. 226:10-16(1991).[3] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt P.C., Kenyon G.L. Biochemistry 29:9856-9862(1990).

644. TPR Domain

30 [1]

Medline: 95397415

Tetratrico peptide repeat interactions: to TPR or not to TPR?

Lamb JR, Tugendreich S, Hieter P;

Trends Biochem Sci 1995;20:257-259.

[2]Medline: 98151343

The structure of the tetratricopeptide repeats of protein
phosphatase 5: implications for TPR-mediated protein-protein
interactions.

Das AK, Cohen PW, Barford D;

EMBO J 1998;17:1192-1199.

Number of members: 621

645. Uroporphyrin-III C-methyltransferase signatures (TP methylase)

Uroporphyrin-III C-methyltransferase (EC 2.1.1.107) (SUMT) [1,2] catalyzes the transfer of
two methyl groups from S-adenosyl-L-methionine to the C-2 and C-7atoms of
uroporphyrinogen III to yield precorrin-2 via the intermediate formation of precorrin-1.

SUMT is the first enzyme specific to the cobalamin pathway and precorrin-2 is a common
intermediate in the biosynthesis of corrinoids such as vitamin B12, siroheme and coenzyme
F430. The sequences of SUMT from a variety of eubacterial and archaeobacterial species are
currently available. In species such as *Bacillus megaterium* (gene *cobA*), *Pseudomonas*
denitrificans (*cobA*) or *Methanobacterium ivanovii* (gene *corA*) SUMT is a protein of about
25 to 30 Kd. In *Escherichia coli* and related bacteria, the *cysG* protein, which is involved in
the biosynthesis of siroheme, is a multifunctional protein composed of a N-terminal domain,
probably involved in transforming precorrin-2 into siroheme, and a C-terminal domain which
has SUMT activity. The sequence of SUMT is related to that of a number of *P. denitrificans*
and *Salmonella typhimurium* enzymes involved in the biosynthesis of cobalamin which also
seem to be SAM-dependent methyltransferases [3,4]. The similarity is especially strong with
two of these enzymes: *cobI/cbiL* which encodes S-adenosyl-L-methionine--precorrin-2
methyltransferase and *cobM/cbiF* whose exact function is not known. Two signature patterns
have been developed for these enzymes. The first corresponds to a well conserved region in
the N-terminal extremity (called region 1 in [1,3]) and the second to a less conserved region
located in the central part of these proteins (this pattern spans what are called regions 2 and 3
in [1,3]).

Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-[GS]-[STAL][STAL SEQ ID NO:471]-
G-P-G-x(3)-[LIVMFY][LIVMFY SEQ ID NO:18]-[LIVM][LIVM SEQ ID NO:4]-T-
[LIVM][LIVM SEQ ID NO:4]-[KRHQG][KRHQG SEQ ID NO:592]-[AG]

Consensus pattern: V-x(2)-[LI]-x(2)-G-D-x(3)-[FYW]-[GS]-x(8)-[LIVF][LIVF SEQ ID
 5 NO:127]-x(5,6)-[LIVMFYWPAC][LIVMFYWPAC SEQ ID NO:593]-x-
[LIVMY][LIVMY SEQ ID NO:141]-x-P-G

[1] Blanche F., Robin C., Couder M., Faucher D., Cauchois L., Cameron B., Crouzet J. J.
 Bacteriol. 173:4637-4645(1991).[2] Robin C., Blanche F., Cauchois L., Cameron B., Couder
 M., Crouzet J. J. Bacteriol. 173:4893-4896(1991).[3] Crouzet J., Cameron B., Cauchois L.,
 10 Rigault S., Rouyez M.-C., Blanche F., Thibaut D., Debussche L. J. Bacteriol. 172:5980-
 5990(1990).[4] Roth J.R., Lawrence J.G., Rubenfield M., Kieffer-Higgins S., Church G.M. J.
 Bacteriol. 175:3303-3316(1993).[5] Mattheakis L.C., Shen W.H., Collier R.J. Mol. Cell.
 Biol. 12:4026-4037(1992).

646. Tudor domain

Domain of unknown function present in several RNA-binding proteins. copies in the
 Drosophila Tudor protein. Slight ambiguities in the alignment.Number of members: 18
 [1]Medline: 97200561 Tudor domains in proteins that interact with RNA. Ponting CP;
 20 Trends Biochem Sci 1997;22:51-52. [2]Medline: 97157029 The human EBNA-2
 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease
 fold and to the tudor protein involved in Drosophila melanogaster development. Callebaut I,
 Mornon JP; Biochem J 1997;321:125-132.

647. Terpene synthase family

It has been suggested that this gene family be designated
 tps (for terpene synthase) [1]. It has been split into six
 subgroups on the basis of phylogeny, called tpsa-tpsf.
 30 tpsa includes vetispiradiene synthase Swiss:Q39979, 5-epi-
 aristolochene synthase, Swiss:Q40577 and (+)-delta-cadinene
 synthase Swiss:P93665.
 tpsb includes (-)-limonene synthase, Swiss:Q40322.

542

tpsc includes kaurene synthase A, Swiss:O04408.

tpsd includes taxadiene synthase, Swiss:Q41594, pinene synthase,
Swiss:O24475 and myrcene synthase, Swiss:O24474.

tpse includes kaurene synthase B.

5 tpsf includes linalool synthase.

Number of members: 51

[1]

Medline: 97413772

10 Monoterpene synthases from grand fir (*Abies grandis*). cDNA
isolation, characterization, and functional expression of
myrcene synthase, (-)-(4S)-limonene synthase, and
(-)-(1S,5S)-pinene synthase.

Bohlmann J, Steele CL, Croteau R;
J Biol Chem 1997;272:21784-21792.

15

648. ThiF family

This family contains a repeated domain in ubiquitin
activating enzyme E1 and members of the bacterial

20 ThiF/MoeB/HesA family. Number of members: 87

649. Thioester dehydrase

Members of this family are involved in fatty acid biosynthesis.

25 Number of members: 19

[1]

Medline: 96398612

30 Structure of a dehydratase-isomerase from the bacterial
pathway for biosynthesis of unsaturated fatty acids: two
catalytic activities in one active site.

Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL;
Structure 1996;4:253-264.

Database Reference: SCOP; 1mka; fa; [SCOP-USA][CATH-PDBSUM]

Database reference: PFAMB; PB058036;

650. Tub family signatures

5 The mouse tubby mutation is the cause of maturity-onset obesity, insulin resistance and sensory deficits. This mutation maps to a gene, *tub* [1,2], which codes for a protein that belongs to a family which currently consists of the following members: - Mammalian *tub*, an hydrophilic protein of about 500 residues, which could be involved in the hypothalamic regulation of body weight. - Human protein TULP1 [3] which may be involved in retinis pigmentosa 14, a retinal degeneration disease. - Mouse protein p4-6 whose function is not
10 known. - *Caenorhabditis elegans* hypothetical protein F10B5.4. - Several fragmentary sequences from plants, *Drosophila* and human ESTs. While the N-terminal part of these protein is not conserved in length nor in the sequence, the C-terminal 250 residues are highly conserved. Therefore, two regions were selected in the C-terminal part as signature patterns.

15 The second region is located at the C-terminal extremity and contains a penultimate cysteine residue that could be critical to the normal functioning of these proteins.

Consensus pattern: F-[KHQ]-G-R-V-[ST]-x-A-S-V-K-N-F-Q

Consensus pattern: A-F-[AG]-I-[SAC]-[LIVM][LIVM SEQ ID NO:4]-[ST]-S-F-x-[GST]-K-
x-A-C-E

20 [1] Kleyn P.W., Fan W., Kovats S.G., Lee J.L., Pulido J.C., Wu Y., Berkemeier L.R., Misumi D.J., Holmgren L., Charlat O., Woolf E.A., Tayber O., Brody T., Shu P., Hawkins F., Kennedy B., Baldini L., Ebeling C., Alperin G.D., Deeds J., Lakey N.D., Culpepper J., Chen H., Gluecksmann-Kuis M.A., Carlson G.A., Duyk G.M., Moore K.J. *Cell* 85:281-290(1996).[2] Noben-Trauth K., Naggert J.K., North M.A., Nishina P.M. *Nature* 380:534-538(1996).[3]
25 North M.A., Naggert J.K., Yan Y., Noben-Trauth K., Nishina P.M. *Proc. Natl. Acad. Sci. U.S.A.* 94:3128-3133(1997).

651. Eukaryotic DNA topoisomerase I active site

30 DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type I topoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a eukaryotic type I topoisomerase breaks a DNA backbone bond, it

simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 3'-phosphate on DNA, at one end of the enzyme-severed DNA strand. In eukaryotes and pox virus topoisomerases I, there are a number of conserved residues in the region around the active site tyrosine.

5 Consensus pattern: [DEN]-x(6)-[GS]-[IT]-S-K-x(2)-Y-[LIVM][LIVM SEQ ID NO:4)]-x(3)-[LIVM][LIVM SEQ ID NO:4)] [Y is the active site tyrosine]

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).[2] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).[3] Lynn R.M., Bjornsti M.-A., Caron P.R., Wang J.C. Proc. Natl. Acad. Sci. U.S.A. 86:3559-3563(1989).[4] Roca J. Trends Biochem. Sci. 10 20:156-160(1995).[E1]

652. Transaldolase signatures

Transaldolase (EC 2.2.1.2) catalyzes the reversible transfer of a three-carbonketol unit from sedoheptulose 7-phosphate to glyceraldehyde 3-phosphate to form erythrose 4-phosphate and fructose 6-phosphate. This enzyme, together with transketolase, provides a link between the glycolytic and pentose-phosphate pathways. Transaldolase is an enzyme of about 34 Kd whose sequence has been well conserved throughout evolution. A lysine has been implicated [1]in the catalytic mechanism of the enzyme; it acts as a nucleophilic group that attacks the carbonyl group of fructose-6-phosphate. Transaldolase is evolutionary related [2] to a bacterial protein of about 20Kd (known as talC in Escherichia coli), whose exact function is not yet known. Two signature patterns have been developed for these proteins. The first, located in the N-terminal section, contains a perfectly conserved pentapeptide; these cond, includes the active site lysine.

25 Consensus pattern: [DG]-[IVSA][IVSA SEQ ID NO:594)]-T-[ST]-N-P-[STA]-[LIVMF][LIVMF SEQ ID NO:2)](2)

Consensus pattern: [LIVM][LIVM SEQ ID NO:4)]-x-[LIVM][LIVM SEQ ID NO:4)]-K-[LIVM][LIVM SEQ ID NO:4)]-[PAS]-x-[ST]-x-[DENQPAS][DENQPAS SEQ ID NO:595)]-G-[LIVM][LIVM SEQ ID NO:4)]-x-[AGV]-x-[QEKRST][QEKRST SEQ ID NO:596)]-x-[LIVM][LIVM SEQ ID NO:4)] [K is the active site residue]

[1] Miosga T., Schaaff-Gerstenschlaeger I., Franken E., Zimmermann F.K. Yeast 9:1241-1249(1993).[2] Reizer J., Reizer A., Saier M.H. Jr. Microbiology 141:961-971(1995).

653. (Transpeptidase) Penicillin binding protein transpeptidase domain

The active site serine (residue 337 in [Swiss:P14677](#)) is conserved in all members of this family.

[1] Pares S, Mouz N, Petillot Y, Hakenbeck R, Dideberg O *Nat Struct Biol* 1996;3:284-289.

654. Trehalase signatures

Trehalase (EC [3.2.1.28](#)) is the enzyme responsible for the degradation of the disaccharide alpha, alpha-trehalose yielding two glucose subunits [1]. It is an enzyme found in a wide variety of organisms and whose sequence has been highly conserved throughout evolution. Two of the most highly conserved regions have been selected as signature patterns. The first pattern is located in the central section, the second one is in the C-terminal region.

Consensus pattern: P-G-G-R-F-x-E-x-Y-x-W-D-x-Y

Consensus pattern: Q-W-D-x-P-x-[GA]-W-[PAS]-P

[1] Kopp M., Mueller H., Holzer H. *J. Biol. Chem.* 268:4766-4774(1993).[2] Henrissat B., Bairoch A. *Biochem. J.* 293:781-788(1993).[E1]

655. Trehalose-6-phosphate synthase domain

OtsA (Trehalose-6-phosphate synthase) is homologous to regions in the subunits of yeast trehalose-6-phosphate synthase/phosphate complex, [1].

[1] Kaasen I, McDougall J, Strom AR; *Gene* 1994;145:9-15.

656. Tropomyosins signature

Tropomyosins [1,2] are family of closely related proteins present in muscle and non-muscle cells. In striated muscle, tropomyosin mediate the interactions between the troponin complex and actin so as to regulate muscle contraction. The role of tropomyosin in smooth muscle and non-muscle tissues is not clear. Tropomyosin is an alpha-helical protein that forms a coiled-coil dimer. Muscle isoforms of tropomyosin are characterized by having 284 amino acid

residues and a highly conserved N-terminal region, whereas non-muscle forms are generally smaller and are heterogeneous in their N-terminal region. The signature pattern for tropomyosins is based on a very conserved region in the C-terminal section of tropomyosins and which is present in both muscle and non-muscle forms.

5 Consensus pattern: L-K-E-A-E-x-R-A-E

[1] Smilie L.B. Trends Biochem. Sci. 4:151-155(1979).[2] McLeod A.R. BioEssays 6:208-212(1986).

10 657. Troponin

Troponin (Tn) contains three subunits, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT). this Pfam contains members of the TnT subunit.

15 Troponin is a complex of three proteins, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT).

The troponin complex regulates Ca⁺⁺ induced muscle contraction.

This family includes troponin T and troponin I. Troponin I binds to actin and troponin T binds to tropomyosin.

Number of members: 81 [1]

20 Medline: 87144593

Structure of co-crystals of tropomyosin and troponin.

White SP, Cohen C, Phillips GN Jr;

Nature 1987;325:826-828. [2]

Medline: 95155315

25 A direct regulatory role for troponin T and a dual role for troponin C in the Ca²⁺ regulation of muscle contraction.

Potter JD, Sheng Z, Pan BS, Zhao J;

J Biol Chem 1995;270:2557-2562.

[3]Medline: 95324796

30 The troponin complex and regulation of muscle contraction.

Farah CS, Reinach FC;

FASEB J 1995;9:755-767.

658. (Tryp mucin) Mucin-like glycoprotein

This family of trypanosomal proteins resemble vertebrate mucins. The protein consists of three regions. The N and C termini are conserved between all members of the family, whereas the central region is not well conserved and contains a large number of threonine residues which can be glycosylated [1].

Indirect evidence suggested that these genes might encode the core protein of parasite mucins, glycoproteins that were proposed to be involved in the interaction with, and invasion of, mammalian host cells.

[1] Di Noia JM, Sanchez DO, Frasch AC; J Biol Chem 1995;270:24146-24149.

[2] Di Noia JM, D'Orso I, Aslund L, Sanchez DO, Frasch AC; J Biol Chem 1998;273:10843-10850.

659. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site. The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.

Consensus pattern: P-x(0,2)-[GSTAN][GSTAN SEQ ID NO:296)]-[DENQGAPK][DENQGAPK SEQ ID NO:597)]-x-[LIVMEFP][LIVMEFP SEQ ID NO:598)]-

[HT]-[LIVMYAC][LIVMYAC SEQ ID NO:599]-G-[HNTG][HNTG SEQ ID NO:600]-
[LIVMFYSTAGPC][LIVMFYSTAGPC SEQ ID NO:601]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M.,
Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow
5 D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-
687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle
R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

660. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino
acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In
prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA
synthetases, one for each different amino acid. In eukaryotes there are generally two
15 aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a
mitochondrial form. While all these enzymes have a common function, they are widely
diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2]
that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal
section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well
20 conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.
The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific
for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine,
tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I
synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.

25 Consensus pattern: P-x(0,2)-[GSTAN][GSTAN SEQ ID NO:296]-
[DENQGAPK][DENQGAPK SEQ ID NO:597]-x-[LIVMFP][LIVMFP SEQ ID NO:598]-
[HT]-[LIVMYAC][LIVMYAC SEQ ID NO:599]-G-[HNTG][HNTG SEQ ID NO:600]-
[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M.,
30 Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow
D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-
687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle
R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

661. (tRNA-synt 1C) tRNA synthetases class I (E and Q)

5 Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only glutamyl and glutaminyl tRNA synthetases.

In some organisms, a single glutamyl-tRNA synthetase aminoacylates both tRNA(Glu) and tRNA(Gln).

10 [1] Rath VL, Silvian LF, Beijer B, Sproat BS, Steitz TA; Structure 1998;6:439-449.

662. (tRNA-synt 1d) tRNA synthetases class I (R)

15 Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only arginyl tRNA synthetase.

663. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2)

20 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a

25 mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is

30 different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF][GSTALVF SEQ ID NO:42)]-(DENQHRRKP)[DENQHRRKP
SEQ ID NO:43)]-[GSTA][GSTA SEQ ID NO:19)]-[LIVMF][LIVMF SEQ ID NO:2)]-[DE]-
R-[LIVMF][LIVMF SEQ ID NO:2)]-x-[LIVMSTAG][LIVMSTAG SEQ ID NO:44)]-
[LIVMFY][LIVMFY SEQ ID NO:18)]

- 5 [1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D.
 BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel
 G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S.,
 Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S.
 Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar
 10 N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet
 S. Nucleic Acids Res. 18:305-312(1990).

664. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1e)

- 15 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino
 acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In
 prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA
 synthetases, one for each different amino acid. In eukaryotes there are generally two
 aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a
 20 mitochondrial form. While all these enzymes have a common function, they are widely
 diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2]
 that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal
 section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well
 conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.
 25 The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific
 for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine,
 tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I
 synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.
 Consensus pattern: P-x(0,2)-[GSTAN][GSTAN SEQ ID NO:296)]-
 30 [DENQGAPK][DENQGAPK SEQ ID NO:597)]-x-[LIVMEP][LIVMEP SEQ ID NO:598)]-
[HT]-[LIVMYAC][LIVMYAC SEQ ID NO:599)]-G-[HNTG][HNTG SEQ ID NO:600)]-
[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

665. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: {GSTALVF}[GSTALVF SEQ ID NO:42]-{DENQHRKP}[DENQHRKP SEQ ID NO:43]-[GSTA][GSTA SEQ ID NO:19]-[LIVMF][LIVMF SEQ ID NO:2]-[DE]-R-[LIVMF][LIVMF SEQ ID NO:2]-x-[LIVMSTAG][LIVMSTAG SEQ ID NO:44]-[LIVMEY][LIVMEY SEQ ID NO:18]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

666. Thaumatin family signature

Thaumatococcus daniellii, an African brush. The protein is made of about 200 residues and contains 8 disulfide bonds. A number of proteins have been found to be related to thaumatins. These protein are listed below (references are only provided for recently determined sequences). - A maize alpha-amylase/trypsin inhibitor. - Two tobacco pathogenesis-related proteins: PR-R major and minor forms, which are induced after infection with viruses. - Salt-induced protein NP24 from tomato. - Osmotin, a salt-induced protein from tobacco. - Osmotin-like proteins OSML13, OSML15 and OSML81 from potato [2]. - P21, a leaf protein from soybean. - PWIR2, a leaf protein from wheat. - Zeamatin, a maize antifungal protein [3]. The exact biological function of all these proteins is not yet known. A conserved region that includes three cysteine residues known (in thaumatocin) to be involved in disulfide bonds has been selected as a signature pattern.

```
+-----+ | +-----+ | | ***** |
||
```

```
xxCxxxxxxxxxxxxxxxxCxxCxxCxxxxxxxxxxxxxxxxCxxCxCxxxCxCxxCCxCxxxCxxxxxC
xxxCx ||||| ||| ++ ++ | +--+ +--+ | +-----+'C': conserved cysteine
```

involved in a disulfide bond. '*' : position of the pattern.

Consensus pattern: G-x-[GF]-x-C-x-T-[GA]-D-C-x(1,2)-G-x(2,3)-C

[1] Edens L., Heslinga L., Klok R., Ledebor A.M., Maat J., Toonen M.Y., Visser C., Verrips C.T. Gene 18:1-12(1982).[2] Zhu B., Chen T.H.H., Li P.H. Plant Physiol. 108:929-937(1995).[3] Malehorn D.E., Borgmeyer J.R., Smith C.E., Shah D.M.; Plant Physiol. 106:1471-1481(1994).

667. Thiolases signatures

Two different types of thiolase [1,2,3] are found both in eukaryotes and in prokaryotes: acetoacetyl-CoA thiolase (EC 2.3.1.9) and 3-ketoacyl-CoA thiolase (EC 2.3.1.16). 3-ketoacyl-CoA thiolase (also called thiolase I) has a broad chain-length specificity for its substrates and is involved in degradative pathways such as fatty acid beta-oxidation. Acetoacetyl-CoA thiolase (also called thiolase II) is specific for the thiolysis of acetoacetyl-CoA and involved

in biosynthetic pathways such as poly beta-hydroxybutyrate synthesis or steroid biogenesis. In eukaryotes, there are two forms of 3-ketoacyl-CoA thiolase: one located in the mitochondrion and the other in peroxisomes. There are two conserved cysteine residues important for thiolase activity. The first located in the N-terminal section of the enzymes is involved in the formation of an acyl-enzyme intermediate; the second located at the C-terminal extremity is the active site base involved in deprotonation in the condensation reaction. Mammalian nonspecific lipid-transfer protein (nsL-TP) (also known as sterol carrier protein 2) is a protein which seems to exist in two different forms: a 14 Kd protein (SCP-2) and a larger 58 Kd protein (SCP-x). The former is found in the cytoplasm or the mitochondria and is involved in lipid transport; the latter is found in peroxisomes. The C-terminal part of SCP-x is identical to SCP-2 while the N-terminal portion is evolutionary related to thiolases[4]. Three signature patterns have been developed for this family of proteins, two of which are based on the regions around the biologically important cysteines. The third is based on a highly conserved region in the C-terminal part of these proteins.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4)]-[NST]-x(2)-C-[SAGLI][SAGLI SEQ ID NO:602)]-[ST]-[SAG]-[LIVMFYNS][LIVMFYNS SEQ ID NO:603)]-x-[STAG][STAG SEQ ID NO:20)]-[LIVM][LIVM SEQ ID NO:4)]-x(6)-[LIVM][LIVM SEQ ID NO:4)] [C is involved in formation of acyl-enzyme intermediate]

Consensus pattern: N-x(2)-G-G-x-[LIVM][LIVM SEQ ID NO:4)]-[SA]-x-G-H-P-x-[GA]-x-[ST]-G

Consensus pattern: [AG]-[LIVMA][LIVMA SEQ ID NO:30)]-[STAGCLIVM][STAGCLIVM SEQ ID NO:604)]-[STAG][STAG SEQ ID NO:20)]-[LIVMA][LIVMA SEQ ID NO:30)]-C-x-[AG]-x-[AG]-x-[AG]-x-[SAG] [C is the active site residue]

[1] Peoples O.P., Sinskey A.J. J. Biol. Chem. 264:15293-15297(1989).[2] Yang S.-Y., Yang X.-Y.H., Healy-Louie G., Schulz H., Elzinga M. J. Biol. Chem. 265:10424-10429(1990).[3] Igual J.C., Gonzalez-Bosch C., Dopazo J., Perez-Ortin J.E. J. Mol. Evol. 35:147-155(1992).[4] Baker M.E., Billheimer J.T., Strauss J.F. III DNA Cell Biol. 10:695-698(1991).

668. Thioredoxin family active site

Thioredoxins [1 to 4] are small proteins of approximately one hundred amino-acid residues which participate in various redox reactions via the reversible oxidation of an active center

disulfide bond. They exist in either a reduced form or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond. Thioredoxin is present in prokaryotes and eukaryotes and the sequence around the redox-active disulfide bond is wellconserved. Bacteriophage T4 also encodes for a thioredoxin but its primary structure is not homologous to bacterial, plant and vertebrate thioredoxins. A number of eukaryotic proteins contain domains evolutionary related to thioredoxin, all of them seem to be protein disulphide isomerases (PDI). PDI(EC 5.3.4.1) [5,6,7] is an endoplasmic reticulum enzyme that catalyzes the rearrangement of disulfide bonds in various proteins. The various forms of PDI which are currently known are: - PDI major isozyme; a multifunctional protein that also function as the beta subunit of prolyl 4-hydroxylase (EC 1.14.11.2), as a component of oligosaccharyl transferase (EC 2.4.1.119), as thyroxine deiodinase (EC 3.8. 1.4), as glutathione-insulin transhydrogenase (EC 1.8.4.2) and as a thyroid hormone-binding protein ! - ERp60 (ER-60; 58 Kd microsomal protein). ERp60 was originally thought to be a phosphoinositide-specific phospholipase C isozyme and later to be a protease. - ERp72. - P5. All PDI contains two or three (ERp72) copies of the thioredoxin domain. Bacterial proteins that act as thiol:disulfide interchange proteins that allows disulfide bond formation in some periplasmic proteins also contain a thioredoxin domain. These proteins are: - Escherichia coli dsbA (or prfA) and its orthologs in Vibrio cholerae (tcpG) and Haemophilus influenzae (por). - Escherichia coli dsbC (or xprA) and its orthologs in Erwinia chrysanthemi and Haemophilus influenzae. - Escherichia coli dsbD (or dipZ) and its Haemophilus influenzae ortholog. - Escherichia coli dsbE (or ccmG) and orthologs in Haemophilus influenzae, Rhodobacter capsulatus (helX), Rhizobiaceae (cycY and tlpA).

Consensus pattern: [LIVMF][LIVMF SEQ ID NO:2]-[LIVMSTA][LIVMSTA SEQ ID NO:433]-x-[LIVMFYC][LIVMFYC SEQ ID NO:6]-[FYWSTHE][FYWSTHE SEQ ID NO:605]-x(2)-[FYWGTN][FYWGTN SEQ ID NO:606]-C-[GATPLVE][GATPLVE SEQ ID NO:607]-[PHYWSTA][PHYWSTA SEQ ID NO:608]-C-x(6)-[LIVMFYWT][LIVMFYWT SEQ ID NO:47] [The two C's form the redox-active bond]

[1] Holmgren A. Annu. Rev. Biochem. 54:237-271(1985).[2] Gleason F.K., Holmgren A. FEMS Microbiol. Rev. 54:271-297(1988).[3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989).[4] Eklund H., Gleason F.K., Holmgren A. Proteins 11:13-28(1991).[5] Freedman R.B., Hawkins H.C., Murrant S.J., Reid L. Biochem. Soc. Trans. 16:96-99(1988).[6] Kivirikko K.I., Myllyla R., Pihlajaniemi T. FASEB J. 3:1609-1617(1989).[7] Freedman R.B., Hirst T.R., Tuite M.F. Trends Biochem. Sci. 19:331-336(1994).

669. (Transcript fac2) Transcription factor TFIIB repeat signature

In eukaryotes the initiation of transcription of protein encoding genes by polymerase II is modulated by general and specific transcription factors. The general transcription factors operate through common promoters elements (such as the TATA box). At least seven different proteins associates to form the general transcription factors: TFIIA, -IIB, -IID, -IIE, -IIF, -IIG, and -IIH[1]. Transcription factor IIB (TFIIB) plays a central role in the transcription of class II genes, it associates with a complex of TFIID-IIA bound to DNA (DA complex) to form a ternary complex TFIID-IIA-IBB (DAB complex) which is then recognized by RNA polymerase II [2,3]. TFIIB is a protein of about 315 to 340 amino acid residues which contains, in its C-terminal part an imperfect repeat of a domain of about 75 residues. This repeat could contribute an element of symmetry to the folded protein. The following proteins have been shown to be evolutionary related to TFIIB: - An archaeobacterial TFIIB homolog. In *Pyrococcus woesei* a previously undetected open reading frame has been shown [4] to be highly related to TFIIB. - Fungal transcription factor IIIB 70 Kd subunit (gene PCF4/TDS4/BRF1) [5]. This protein is a general activator of RNA polymerase III transcription and plays a role analogous to that of TFIIB in pol III transcription. The central section of the repeated domain, which is the most conserved part of that domain has been selected as a signature pattern.

Consensus pattern: G-[KR]-x(3)-[STAGN][STAGN SEQ ID NO:24])-x-[LIVMYA][LIVMYA SEQ ID NO:609])- [GSTA][GSTA SEQ ID NO:19])(2)-[CSAV][CSAV SEQ ID NO:155])- [LIVM][LIVM SEQ ID NO:4])- [LIVMFY][LIVMFY SEQ ID NO:18])- [LIVMA][LIVMA SEQ ID NO:30])- [GSA]-[STAC

[1] Weinmann R. Gene Expr. 2:81-91(1992).[2] Hawley D. Trends Biochem. Sci. 16:317-318(1991).[3] Ha I., Lane W.S., Reinberg D. Nature 352:689-695(1991).[4] Ouzounis C., Sander C. Cell 71:189-190(1992).[5] Khoo B., Brophy B., Jackson S.P. Genes Dev. 8:2879-2890(1994).

670. (transcript fact) MADS-box domain signature and profile

A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below: - Serum response

factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1. - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamina into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6]. - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). In SRF, the conserved domain has been shown [1] to be involved in DNA-binding and dimerization. A pattern that spans the complete length of the domain has been derived. The profile also spans the length of the MADS-box.

Consensus pattern: R-x-[RK]-x(5)-I-x-[DNGSK][DNGSK SEQ ID NO:610]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM][LIVM SEQ ID NO:4]-x-K(2)-A-x-E-[LIVM][LIVM SEQ ID NO:4]-[STA]-x-L-x(4)-[LIVM][LIVM SEQ ID NO:4]-x-[LIVM][LIVM SEQ ID NO:4](3)-x(6)-[LIVMF][LIVMF SEQ ID NO:2]-x(2)-[FY]

[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).[2]

Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).[3]

Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature

346:35-39(1990).[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).[5]

Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer

H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992).[6] Ma H., Yanofsky M.F.,

Meyerowitz E.M. Genes Dev. 5:484-495(1991).[E1]

671. Transketolase signatures

Transketolase (EC 2.2.1.1) (TK) catalyzes the reversible transfer of a two-carbon ketol unit from xylulose 5-phosphate to an aldose receptor, such as ribose 5-phosphate, to form sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate. This enzyme, together with transaldolase, provides a link between the glycolytic and pentose-phosphate pathways. TK requires thiamin pyrophosphate as a cofactor. In most sources where TK has been purified, it is a homodimer of approximately 70 Kd subunits. TK sequences from a variety of eukaryotic and prokaryotic sources [1,2] show that the enzyme has been evolutionarily conserved. In the peroxisomes of methylotrophic yeast *Hansenula polymorpha*, there is a highly related enzyme, dihydroxy-acetone synthase (DHAS) (EC 2.2.1.3) (also known as formaldehyde transketolase), which exhibits a very unusual specificity by including formaldehyde amongst its substrates. 1-deoxyxylulose-5-phosphate synthase (DXP synthase) [3] is an enzyme so far found in bacteria (gene *dxs*) and plants (gene *CLA1*) which catalyzes the thiamin pyrophosphate-dependent acyloin condensation reaction between carbon atoms 2 and 3 of pyruvate and glyceraldehyde 3-phosphate to yield 1-deoxy-D- xylulose-5-phosphate (dxp), a precursor in the biosynthetic pathway to isoprenoids, thiamin (vitamin B1), and pyridoxol (vitamin B6). DXP synthase is evolutionary related to TK. Two regions of TK have been selected as signature patterns. The first, located in the N-terminal section, contains a histidine residue which appears to function in proton transfer during catalysis [4]. The second, located in the central section, contains conserved acidic residues that are part of the active cleft and may participate in substrate-binding [4].

Consensus pattern: R-x(3)-[LIVMTA][LIVMTA SEQ ID NO:311]-[DENQSTHKF][DENQSTHKF SEQ ID NO:611]-x(5,6)-[GSN]-G-H-[PLIVMF][PLIVMF SEQ ID NO:612]-[GSTA][GSTA SEQ ID NO:19]-x(2)-[LIMC][LIMC SEQ ID NO:613]-[GS

Consensus pattern: G-[DEQGSA][DEQGSA SEQ ID NO:614]-[DN]-G-[PAEQ][PAEQ SEQ ID NO:615]-[ST]-[HQ]-x-[PAGM][PAGM SEQ ID NO:616]-[LIVMYAC][LIVMYAC SEQ ID NO:599]-[DEFYW][DEFYW SEQ ID NO:617]-x(2)-[STAP][STAP SEQ ID NO:135]-x(2)-[RGA]

[1] Abedinia M., Layfield R., Jones S.M., Nixon P.F., Mattick J.S. *Biochem. Biophys. Res. Commun.* 183:1159-1166(1992).[2] Fletcher T.S., Kwee I.L., Nakada T., Largman C., Martin B.M. *Biochemistry* 31:1892-1896(1992).[3] Sprenger G.A., Schorken U., Wiegert T.,

A conserved region that includes two cysteines and seems to be located in a short cytoplasmic loop between two transmembrane domains has been selected as a signature for these proteins.

Consensus pattern: G-x(3)-[LIVMF][LIVMF SEQ ID NO:2)]-x(2)-[GSA]-[LIVMF][LIVMF
5 SEQ ID NO:2)](2)-G-C-x-[GA]-[STA]-x(2)-[EG]-x(2)-[CWN]-[LIVM][LIVM SEQ ID
NO:4)](2)

[1] Levy S., Nguyen V.Q., Andria M.L., Takahashi S. J. Biol. Chem. 266:14597-
14602(1991).[2] Tomlinson M.G., Williams A.F., Wright M.D. Eur. J. Immunol. 23:136-
40(1993).[3] Barclay A.N., Birkeland M.L., Brown M.H., Beyers A.D., Davis S.J., Somoza
10 C., Williams A.F. The leucocyte antigen factbooks. Academic Press, London / San Diego,
(1993).

673. Tryptophan synthase alpha chain signature

15 Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion
of indoleglycerol phosphate and serine, to tryptophan and glyceraldehyde 3-phosphate [1,2]. It
has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole
and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and
serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta
20 chains), while in fungi the two domains are fused together on a single multifunctional protein.
A conserved region that contains three conserved acidic residues has been selected as a
signature pattern for the alpha chain. The first and the third acidic residues are believed to
serve as proton donors/acceptors in the enzyme's catalytic mechanism.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4)]-E-[LIVM][LIVM SEQ ID NO:4)]-G-
25 x(2)-[FYC]-[ST]-[DE]-[PA]-[LIVMY][LIVMY SEQ ID NO:141)]-[AGLH][AGLH SEQ ID
NO:618)]-[DE]-G

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W.
Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci.
U.S.A. 86:4604-4608(1989).

674. Tryptophan synthase beta chain pyridoxal-phosphate attachment site

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, to tryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. The beta chain of the enzyme requires pyridoxal-phosphate as a cofactor. The pyridoxal-phosphate group is attached to a lysine residue. The region around this lysine residue also contains two histidine residues which are part of the pyridoxal-phosphate binding site. The signature pattern for the tryptophan synthase beta chain is derived from that conserved region.

-Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-x-H-x-G-[STA]-H-K-x-N [K is the pyridoxal-P attachment site]

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

675. Serine proteases, trypsin family, active sites

The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below. - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. -

Chymotrypsins. - Complement components C1r, C1s, C2, and complement factors B, D and I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin). - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin. - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). - Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type, and tissue-type). - Trypsins I, II, III, and IV. - Trypsases. - Snake venom proteases such as ancrod, batroxobin,

cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab. - Apolipoprotein(a). - Blood fluke cercarial protease. - Drosophila trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. All the above proteins belong to family S1 in the classification of peptidases[2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar enough in the regions of the active site residues that they can be picked up by the same patterns. These proteases are listed below. - Achromobacter lyticus protease I. - Lysobacter alpha-lytic protease. - Streptogrisin A and B (Streptomyces proteases A and B). - Streptomyces griseus glutamyl endopeptidase II. - Streptomyces fradiae proteases 1 and 2. Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-[ST]-A-[STAG][STAG SEQ ID NO:20]-H-C [H is the active site residue] Consensus pattern: [DNSTAGC][DNSTAGC SEQ ID NO:619]-[GSTAPIMVQH][GSTAPIMVQH SEQ ID NO:620]-x(2)-G-[DE]-S-G-[GS]-[SAPHV][SAPHV SEQ ID NO:621]-[LIVMFYWH][LIVMFYWH SEQ ID NO:591]-[LIVMFYSTANQH][LIVMFYSTANQH SEQ ID NO:622] [S is the active site residue] [1] Brenner S. Nature 334:528-530(1988).[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

676. (tsp) Thrombospondin type 1 domain

[1] Bork P; FEBS lett 1993;327:125-130.

677. Tubulin subunits alpha, beta, and gamma signature

Tubulins [1,2], the major constituent of microtubules are dimeric proteins which consist of two closely related subunits (alpha and beta). Tubulin binds two molecules of GTP at two different sites (N and E). At the E (Exchangeable) site, GTP is hydrolyzed during incorporation into the microtubule. Near the E site is an invariant region rich in glycines which is found in both chains and which is now [3] said to control the access of the nucleotide to its binding site. A signature pattern was developed from this region. With the exception of the simple eukaryotes, most species express a variety of closely related alpha and beta

isotypes. In most species there is a third member of the tubulin family: gamma tubulin. Gamma tubulin is found at microtubule organizing centers (MTOC) such as the spindle poles or the centrosome, suggesting that it is involved in the minus-end nucleation of microtubule assembly [4].

5 Consensus pattern: [SAG]-G-G-T-G-[SA]-G

[1] Cleveland D.W., Sullivan K.F. Annu. Rev. Biochem. 54:331-365(1985).[2] Joshi H.C., Cleveland D.W. Cell Motil. Cytoskeleton 16:159-163(1990).[3] Hesse J., Thierauf M., Ponstingl H. J. Biol. Chem. 262:15472-15475(1987).[4] Joshi H.C. BioEssays 15:637-643(1993).

10

Tubulin-beta mRNA autoregulation signal

The stability of beta-tubulin mRNAs are autoregulated by their own translation product [1]. Unpolymerized tubulin subunits bind directly (or activate a factor(s) which binds co-translationally) to the nascent N-terminus of beta-tubulin. This binding is transduced through the adjacent ribosomes to activate an RNase that degrades the polysome-bound mRNA. The recognition element has been shown to be the first four amino acids of beta-tubulin: Met-Arg-Glu-Ile. Mutations to this sequence abolish the autoregulation effect (except for the replacement of Glu by Asp); transposition of this sequence to an internal region of a polypeptide also suppresses the autoregulatory effect.

15

20 Consensus pattern: <M-R-[DE]-[IL]

[1] Cleveland D.W. Trends Biochem. Sci. 13:339-343(1988).

678. (tRNA-synt 2c) Aminoacyl-transfer RNA synthetases class-II signatures. Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common

25

30

folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

5

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]-

Consensus pattern: [GSTALVF][GSTALVF SEQ ID NO:42]-[DENQHRKP][DENQHRKP
SEQ ID NO:43]-[GSTA][GSTA SEQ ID NO:19]-[LIVMF][LIVMF SEQ ID NO:2]-[DE]-
R-[LIVMF][LIVMF SEQ ID NO:2]-x-[LIVMSTAG][LIVMSTAG SEQ ID NO:44]-
[LIVMEY][LIVMEY SEQ ID NO:18]-

10

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S.,
Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S.
Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar
N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet
S. Nucleic Acids Res. 18:305-312(1990).

15

20

679. UBA-domain

The UBA-domain (ubiquitin associated domain) is a novel sequence motif found in several proteins having connections to ubiquitin and the ubiquitination pathway. The structure of the UBA domain consists of a compact three helix bundle [1]. Number of
members: 84

25

[1] Structure of a human DNA repair protein UBA domain that interacts with HIV-1 Vpr. Dieckmann T, Withers-Ward ES, Jarosinski MA, Liu CF, Chen IS, Feigon J; Nat Struct Biol 1998;5:1042-1047.

30

680. UBX domain

Domain present in ubiquitin-regulatory proteins. Present in FAF1 and Shp1p. Number of
members: 19

[1] The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. Hofmann K, Bucher P; Trends Biochem Sci 1996;21:172-173.

- 5 681. (UCH) Ubiquitin carboxyl-terminal hydrolases family 1 cysteine active site
Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The first class consist
- 10 of enzymes of about 25 Kd and is currently represented by: - Mammalian isozymes L1 and L3. - Yeast YUH1. - Drosophila Uch. One of the active site residues of class-I UCH [3] is a cysteine. A signature pattern has been derived from the region around that residue.
Consensus pattern: Q-x(3)-N-[SA]-C-G-x(3)-[LIVM][LIVM SEQ ID NO:4]](2)-H-[SA]-[LIVM][LIVM SEQ ID NO:4]]-[SA] [C is the active site residue
- 15 [1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Johnston S.C., Larsen C.N., Cook W.J., Wilkinson K.D., Hill C.P. EMBO J. 16:3787-3796(1997).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

- 20 682. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-1)
Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well
- 25 as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of large proteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12, UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat
- 30 facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. - Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The

second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY][LIVMFY SEQ ID NO:18]-x(1,3)-[AGC]-[NASM][NASM SEQ ID NO:623]-x-C-[FYW]-[LIVMC][LIVMC SEQ ID NO:142]-[NST]-[SACV][SACV SEQ ID NO:391]-x-[LIVMS][LIVMS SEQ ID NO:429]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT][LIVMFT SEQ ID NO:282]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991).[2] D'andrea A., Pellman D. *Crit. Rev. Biochem. Mol. Biol.* 33:337-352(1998).[3] Rawlings N.D., Barrett A.J. *Meth. Enzymol.* 244:461-486(1994).

683. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-2)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12, UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. -

Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY][LIVMFY SEQ ID NO:18]-x(1,3)-[AGC]-[NASM][NASM SEQ ID NO:623]-x-C-[FYW]-[LIVMC][LIVMC SEQ ID NO:142]-[NST]-[SACV][SACV

SEQ ID NO:391)]-x-[LIVMS][LIVMS SEQ ID NO:429)]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT][LIVMFT SEQ ID NO:282)]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

- 5 [1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991).[2] D'andrea A., Pellman D. *Crit. Rev. Biochem. Mol. Biol.* 33:337-352(1998).[3] Rawlings N.D., Barrett A.J. *Meth. Enzymol.* 244:461-486(1994).

10 684. UDP-glycosyltransferases signature

UDP glycosyltransferases (UGT) are a superfamily of enzymes that catalyzes the addition of the glycosyl group from a UTP-sugar to a small hydrophobic molecule. This family currently consist of: - Mammalian UDP-glucuronosyl transferases (UDPGT) [1,2]. A large family of membrane-bound microsomal enzymes which catalyze the transfer of glucuronic acid to a wide variety of exogenous and endogenous lipophilic substrates. These enzymes are of major importance in the detoxification and subsequent elimination of xenobiotics such as drugs and carcinogens. - A large number of putative UDPGT from *Caenorhabditis elegans*. - Mammalian 2-hydroxyacylsphingosine 1-beta-galactosyltransferase [3] (also known as UDP-galactose-ceramide galactosyltransferase). This enzyme catalyzes the transfer of galactose to ceramide, a key enzymatic step in the biosynthesis of galactocerebrosides, which are abundant sphingolipids of the myelin membrane of the central nervous system and peripheral nervous system. - Plants flavonol O(3)-glucosyltransferase. An enzyme [4] that catalyzes the transfer of glucose from UDP-glucose to a flavanol. This reaction is essential and one of the last steps in anthocyanin pigment biosynthesis. - Baculoviruses ecdysteroid UDP-glucosyltransferase (EC 2.4.1.-) [5] (egt). This enzyme catalyzes the transfer of glucose from UDP-glucose to ectysteroids which are insect molting hormones. The expression of egt in the insect host interferes with the normal insect development by blocking the molting process. - Prokaryotic zeaxanthin glucosyl transferase (gene crtX), an enzyme involved in carotenoid biosynthesis and that catalyses the glycosylation reaction which converts zeaxanthin to zeaxanthin-beta- diglucoside. - *Streptomyces* macrolide glycosyltransferases [6]. These enzymes specifically inactivates macrolide anitibiotics via 2'-O-glycosylation using UDP-glucose. These enzymes share a conserved domain of about 50 amino acid residues located in their C-terminal section and from which a pattern has been extracted to detect them.

567

Consensus pattern: [FW]-x(2)-Q-x(2)-[LIVMYA][LIVMYA SEQ ID NO:609)]-
 [LIMV][LIMV SEQ ID NO:34)]-x(4,6)-[LVGAC][LVGAC SEQ ID NO:624)]-
 [LVFYA][LVFYA SEQ ID NO:625)]- [LIVMF][LIVMF SEQ ID NO:2)]-
 [STAGCM][STAGCM SEQ ID NO:626)]-[HNQ]-[STAGC][STAGC SEQ ID NO:45)]-G-
 5 x(2)-[STAG][STAG SEQ ID NO:20)]-x(3)-[STAGL][STAGL SEQ ID NO:627)]-
 [LIVMFA][LIVMFA SEQ ID NO:81)]-x(4)-[PQR]-[LIVMT][LIVMT SEQ ID NO:1)]-x(3)-
 [PA]-x(3)-[DES]-[QEHN][QEHN SEQ ID NO:628)]

[1] Dutton G.J. (In) Glucoronidation of drugs and other compounds, Dutton G.J., Ed., pp 1-
 78, CRC Press, Boca Raton, (1980).[2] Burchell B., Nebert D.W., Nelson D.R., Bock K.W.,
 10 Iyanagi T., Jansen P.L., Lancet D., Mulder G.J., Chowdhury J.R., Siest G., Tephly T.R.,
 Mackenzie P.I. DNA Cell Biol. 10:487-494(1991).[3] Schulte S., Stoffel W. Proc. Natl.
 Acad. Sci. U.S.A. 90:10265-10269(1993).[4] Furtek D., Schiefelbein J.W., Johnston F.,
 Nelson O.E. Jr. Plant Mol. Biol. 11:473-481(1988).[5] O'Reilly D.R., Miller L.K. Science
 245:1110-1112(1989).[6] Hernandez C., Olano C., Mendez C., Salas J.A. Gene 134:139-
 15 140(1993).

685. UDP-glucose/GDP-mannose dehydrogenase family

The UDP-glucose/GDP-mannose dehydrogenases are a small group of enzymes
 20 which possesses the ability to catalyze the NAD-dependent 2-fold oxidation of an alcohol to
 an acid without the release of an aldehyde intermediate [2]. Number of members: 55

[1] Purification and characterization of guanosine diphospho-D-mannose
 dehydrogenase. A key enzyme in the biosynthesis of alginate by *Pseudomonas aeruginosa*.
 Roychoudhury S, May TB, Gill JF, Singh SK, Feingold DS, Chakrabarty AM; J Biol Chem
 25 1989;264:9380-9385. [2] Properties and kinetic analysis of UDP-glucose dehydrogenase
 from group A streptococci. Irreversible inhibition by UDP-chloroacetol. Campbell RE, Sala
 RF, van de Rijn I, Tanner ME; J Biol Chem 1997;272:3416-3422.

686. Uracil-DNA glycosylase signature

Uracil-DNA glycosylase (EC 3.2.2.-) (UNG) [1] is a DNA repair enzyme that excises uracil
 residues from DNA by cleaving the N-glycosylic bond. Uracil in DNA can arise as a result of
 misincorporation of dUMP residues by DNA polymerase or deamination of cytosine. The

sequence of uracil-DNA glycosylase is extremely well conserved [2] in bacteria and eukaryotes as well as in herpes viruses. More distantly related uracil-DNA glycosylases are also found in poxviruses [3]. In eukaryotic cells, UNG activity is found in both the nucleus and the mitochondria. Human UNG1 protein is transported to both the mitochondria and the nucleus [4]. The N-terminal 77 amino acids of UNG1 seem to be required for mitochondrial localization [4], but the presence of a mitochondrial transitpeptide has not been directly demonstrated. As a signature for this type of enzyme, the most N-terminal conserved region has been selected. This region contains an aspartic acid residue which has been proposed, based on X-ray structures [5,6] to act as a general base in the catalytic mechanism.

Consensus pattern: [KR]-[LIV]-~~[LIV]~~[LIVC SEQ ID NO:629)]-[LIVM][LIVM SEQ ID NO:4)]-x-G-[QI]-D-P-Y [D is the active site residue]-

[1] Sancar A., Sancar G.B. Annu. Rev. Biochem. 57:29-67(1988).[2] Olsen L.C., Aasland R., Wittwer C.U., Krokan H.E., Helland D.E. EMBO J. 8:3121-3125 (1989).[3] Upton C., Stuart D.T., McFadden G. Proc. Natl. Acad. Sci. U.S.A. 90:4518-4522(1993).[4] Slupphaug G., Markussen F.-H., Olsen L.C., Aasland R., Aarsaether N., Bakke O., Krokan H.E., Helland D.E. Nucleic Acids Res. 21:2579-2584(1993).[5] Savva R., McAuley-Hecht K., Brown T., Pearl L. Nature 373:487-493(1995).[6] Mol C.D., Arvai A.S., Slupphaug G., Kavli B., Alseth I., Krokan H.E., Tainer J.A. Cell 80:869-878(1995).[7] Muller S.J., Caradonna S. Biochim. Biophys. Acta 1088:197-207(1991).[8] Meyer-Siegler K., Mauro D.J., Seal G., Wurzer J., Deriel J.K., Sirover M.A. Proc. Natl. Acad. Sci. U.S.A. 88:8460-8464(1991).[9] Muller S.J., Caradonna S. J. Biol. Chem. 268:1310-1319(1993).[10] Barnes D.E., Lindahl T., Sedgwick B. Curr. Opin. Cell Biol. 5:424-433(1993).

687. Uncharacterized protein family UPF0001 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL036c. - *Caenorhabditis elegans* hypothetical protein F09E5.8. - *Bacillus subtilis* hypothetical protein ylmE. - *Escherichia coli* hypothetical protein yggS and HI0090, the corresponding *Haemophilus influenzae* protein. - *Helicobacter pylori* hypothetical protein HP0395. - *Mycobacterium tuberculosis* hypothetical protein MtCY270.20. - *Synechocystis* strain PCC 6803 hypothetical protein slr0556. - *A Pseudomonas aeruginosa* hypothetical protein in pilT 5' region. - *A Vibrio alginolyticus* hypothetical protein in pilT 5' region. These are proteins of from 25 to 30 Kd which contain a

number of conserved regions. The best conserved region which is located in the first third of these proteins has been selected as a signature pattern.

Consensus pattern: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV]

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

5

688. Uncharacterized protein family UPF0003 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli protein aefA. - Escherichia coli hypothetical protein yggB. - Escherichia coli

10 hypothetical protein yjeP and HI0195.1, the corresponding Haemophilus influenzae protein. -

Escherichia coli hypothetical protein ynaI. - Bacillus subtilis hypothetical protein yhdY. -

Helicobacter pylori hypothetical protein HP0415. - Synechocystis strain PCC 6803

hypothetical protein slr0639. - Archaeoglobus fulgidus hypothetical protein AF1546. -

Methanococcus jannaschii hypothetical protein MJ0170. - Methanococcus jannaschii

15 hypothetical protein MJ1143. The size of these proteins range from 30 to 120 Kd. They all

contain a number of transmembrane regions. The best conserved region which is located in and just after the last potential transmembrane region has been selected as a signature

pattern,.

Consensus pattern: G-[STIF][STIF SEQ ID NO:630]-V-x(2)-[LIVM][LIVM SEQ ID

20 NO:4]-x(6)-[LIVMF][LIVMF SEQ ID NO:2]-x(3)-[DQ]-x(3)-[LIV]-x-[LIV]-P-N-x(2)-

[LIVMF][LIVMF SEQ ID NO:2]-[LIVFSTA][LIVFSTA SEQ ID NO:205]-x(5)-N

[1] Bairoch A. Unpublished observations (1997).

25 689. Uncharacterized protein family UPF0004 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli hypothetical protein yliG. - Escherichia coli hypothetical protein yleA and

HI0019, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical

protein yqeV. - Helicobacter pylori hypothetical protein HP0269. - Helicobacter pylori

30 hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA

5'region. - Mycobacterium leprae hypothetical protein B2235_C2_195. - Pseudomonas

aeruginosa hypothetical protein in hemL 3'region. - Synechocystis strain PCC 6803

hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein slI0996. -

Methanococcus jannaschii hypothetical protein MJ0865. - Methanococcus jannaschii hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5. The size of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of which are clustered in a region that can be used as a signature pattern.

5 Consensus pattern: ~~[LIVM]~~[LIVM SEQ ID NO:4)]-x-~~[LIVMT]~~[LIVMT SEQ ID NO:1)]-x(2)-G-C-x(3)-C-~~[STAN]~~[STAN SEQ ID NO:250)]-[FY]-C-x-~~[LIVM]~~[LIVM SEQ ID NO:4)]- x(4)-G

[1] Bairoch A. Unpublished observations (1997).

10

690. Uncharacterized protein family UPF0005 signature

The following proteins seem to be evolutionary related [1]: - Mammalian protein TEGT (Testis Enhanced Gene Transcript). - Escherichia coli hypothetical protein yccA and HI0044, the corresponding Haemophilus influenzae protein. - A probable Pseudomonas aeruginosa ortholog of yccA. These are proteins of about 25 Kd which seem to contain seven transmembrane domains. A signature pattern that corresponds to a region that starts with the beginning of the third transmembrane domain and ends in the middle of the fourth one has been developed.

15

Consensus pattern: G-~~[LIVM]~~[LIVM SEQ ID NO:4)](2)-[SA]-x(5,8)-G-x(2)-~~[LIVM]~~[LIVM SEQ ID NO:4)]-G-P-x-L-x(4)-[SAG]- x(4,6)-~~[LIVM]~~[LIVM SEQ ID NO:4)](2)-x(2)-A-x(3)-T-A-~~[LIVM]~~[LIVM SEQ ID NO:4)](2)-F

20

[1] Walter L., Marynen P., Szpirer J., Levan G., Guenther E. Genomics 28:301-304(1995).

25

691. Uncharacterized protein family UPF0006 signatures

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL055c. - Escherichia coli hypothetical protein ycfH and HI0454, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yigW. - Escherichia coli hypothetical protein yjjV and HI0081, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yabD. - Haemophilus influenzae hypothetical protein HI1664. - Mycoplasma genitalium hypothetical protein MG009. These are proteins of from 24 to 47 Kd which contain a number of conserved regions. They can be picked up in the database by the following patterns.

30

571

Consensus pattern: [LIVMFY][LIVMFY SEQ ID NO:18]](2)-D-[STA]-H-x-H-
[LIVMF][LIVMF SEQ ID NO:2)]-[DN

Consensus pattern: P-[LIVM][LIVM SEQ ID NO:4)]-x-[LIVM][LIVM SEQ ID NO:4)]-H-x-
R-x-[TA]-x-[DE

5 Consensus pattern: [LVSA][LVSA SEQ ID NO:631)]-[LIVA][LIVA SEQ ID NO:219)]-
x(2)-[LIVM][LIVM SEQ ID NO:4)]-[PS]-x(3)-L-[LIVM][LIVM SEQ ID NO:4)]-
[LIVMS][LIVMS SEQ ID NO:429)]-E-T- D-x-P

[1] Bairoch A., Rudd K.E. Unpublished observations (1995).

10

692. Uncharacterized protein family UPF0007 signature

The following proteins seems to be evolutionary related [1]: - Escherichia coli hypothetical protein ygbP and HI0672, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yacM. - Mycobacterium tuberculosis hypothetical protein

15 MtCY06G11.29c. - Synechocystis strain PCC 6803 hypothetical protein slr0951. - A Rhodobacter capsulatus hypothetical protein in nifR3 5'region. Except for the Rhodobacter protein which contains a C-terminal extension, all these proteins have from 225 to 236 amino acids. They are hydrophilic proteins that can be picked up in the database by the following pattern.

20 Consensus pattern: V-L-[IV]-H-D-[GA]-A-R

[1] Bairoch A. Unpublished observations (1997).

693. Uncharacterized protein family UPF0015 signature

25 The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBR002c. - Yeast chromosome XIII hypothetical protein YMR101c. - Escherichia coli hypothetical protein yaeU and HI0920, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP1221. - Mycobacterium leprae hypothetical protein B1937_F2_65. - A Corynebacterium glutamicum hypothetical protein in aroF 3'region. - A Streptomyces fradiae hypothetical protein in transposon Tn4556. - Synechocystis strain PCC 6803 hypothetical protein slI0505. - Methanococcus jannaschii hypothetical protein MJ1372. These are proteins of about 26 to

30

40 Kd whose central region is well conserved. They can be picked up in the database by the following pattern.

Consensus pattern: [DE]-[LIVMF][LIVMF SEQ ID NO:2]](3)-R-T-[SG]-G-x(2)-R-x-S-x-[FY]-[LIVM][LIVM SEQ ID NO:4]](2)-W-Q-

5 [1] Wolfe K.H., Lohan A.J.E. Yeast 10:S41-S46(1994).

694. Uncharacterized protein family UPF0016 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

10 Yeast hypothetical protein YBR187w. - Fission yeast hypothetical protein SpAC17G8.08c. -
 Mouse protein pFT27. - Synechocystis strain PCC 6803 hypothetical protein sll0615. These
 are hydrophobic proteins of 200 to 320 amino acids that seem to contain six or seven
 transmembrane domains. A conserved region which seems, in the eukaryotic proteins of this
 family, to directly follow the second transmembrane domain has been selected as a signature
 15 pattern.

Consensus pattern: E-[LIVM][LIVM SEQ ID NO:4]]-G-D-K-T-F-[LIVMF][LIVMF SEQ ID
 NO:2]](2)-A-

[1] Bairoch A. Unpublished observations (1996).

20

695. Uncharacterized protein family UPF0021 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Yeast chromosome VII hypothetical protein YGL211w. - Dictyostelium discoideum protein
 veg136. - Methanococcus jannaschii hypothetical proteins MJ1157 and MJ1478. These are
 25 proteins of from 300 to 360 residues. They can be picked up in the database by the following
 pattern which is located in their N-terminal section.

Consensus pattern: C-K-x(2)-F-x(4)-E-x(22,23)-S-G-G-K-D

[1] Bairoch A. Unpublished observations (1997).

30

696. Uncharacterized protein family UPF0023 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Mouse protein 22A3. - Yeast chromosome XII hypothetical protein YLR022c. -

573

Caenorhabditis elegans hypothetical protein W06E11.4. - Methanococcus jannaschii hypothetical protein MJ0592. These are hydrophilic proteins of about 30 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: D-x-D-E-[LIV]-L-x(4)-V-F-x(3)-S-K-G-

5 [1] Bairoch A. Unpublished observations (1997).

697. Uncharacterized protein family UPF0024 signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein ygbO and HI0701, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP0926. - Yeast chromosome XV hypothetical protein YOR243c. - Caenorhabditis elegans hypothetical protein B0024.11. - Methanococcus jannaschii hypothetical proteins MJ0588 and MJ1364. These are hydrophilic proteins of from 39 to 77 Kd. They can be picked up in the database by the following pattern.

15 Consensus pattern: G-x-K-D-[KR]-x-A-[LV]-T-x-Q-x-[LIVF][LIVF SEQ ID NO:127]-[SGC]-

[1] Bairoch A. Unpublished observations (1997).

698. Uncharacterized protein family UPF0025 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yfcE. - Bacillus subtilis hypothetical protein ysnB. - Mycoplasma genitalium and pneumoniae hypothetical protein MG207. - Methanococcus jannaschii hypothetical proteins MJ0623 and MJ0936. These are hydrophilic proteins of about 20 Kd. They can be picked up in the database by the following pattern.

25 Consensus pattern: D-V-[LIV]-x(2)-G-H-[ST]-H-x(12)-[LIVMF][LIVMF SEQ ID NO:2]-N-P-G

30 [1] Bairoch A. Unpublished observations (1997).

699. Uncharacterized protein family UPF0029 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome III hypothetical protein YCR59c. - Yeast chromosome IV hypothetical protein YDL177C. - Escherichia coli hypothetical protein yigZ and HI0722, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yvyE. - A Thermus aquaticus hypothetical protein in pol 5'region. These proteins can be picked up in the database by the following pattern.

Consensus pattern: G-x(2)-[LIVM][LIVM SEQ ID NO:4](2)-x(2)-[LIVM][LIVM SEQ ID NO:4]-x(4)-[LIVM][LIVM SEQ ID NO:4]-x(5)-[LIVM][LIVM SEQ ID NO:4](2)-x- R-[FYW](2)-G-G-x(2)-[LIVM][LIVM SEQ ID NO:4]-G

[1] Koonin E.V., Bork P., Sander C. EMBO J. 13:493-503(1994).

700. Uncharacterized protein family UPF0030 signature

The following uncharacterized proteins have been shown [1] to be highly similar: - Yeast chromosome VI hypothetical protein YFL060c. - Yeast chromosome XIII hypothetical protein YMR095c. - Yeast chromosome XIV hypothetical protein YNL334c. - Bacillus subtilis hypothetical protein yaaE. - Haemophilus influenzae hypothetical protein HI1648. - Methanococcus jannaschii hypothetical protein MJ1661. These are hydrophilic proteins of about 19 to 25 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: [GA]-L-I-[LIV]-P-G-G-E-S-T-[STA]

[1] Bairoch A. Unpublished observations (1997).

701. Uncharacterized protein family UPF0032 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yigU and HI0188, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein ycbT. - Mycobacterium tuberculosis hypothetical protein MtCY49.33c and U2126A, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803 hypothetical protein sll0194. - Odontella sinensis and Porphyra purpurea chloroplast hypothetical protein ycf43. These proteins have from 245 to 317 amino acids and seem to contain at least six or seven transmembrane regions. A conserved region located in the central section of these proteins has been developed as a signature pattern.

575

Consensus pattern: Y-x(2)-F-[LIVMA][LIVMA SEQ ID NO:30](2)-x-L-x(4)-G-x(2)-F-[EQ]-[LIVMF][LIVMF SEQ ID NO:2]-P-[LIVM][LIVM SEQ ID NO:4] -

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

5

702. Uncharacterized protein family UPF0034 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli hypothetical protein yhdG and HI0979, the corresponding Haemophilus

influenzae protein. - Escherichia coli hypothetical protein yjbN and HI0634, the

10 corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yohI

and HI0270, the corresponding Haemophilus influenzae protein. - Bacillus subtilis

hypothetical protein yacF. - Rhodobacter capsulatus protein nifR3 and related proteins in

Azospirillum brasilense and Rhizobium leguminosarum. - Synechocystis strain PCC 6803

hypothetical protein slr0644. - Synechocystis strain PCC 6803 hypothetical protein slI0926. -

15 Caenorhabditis elegans hypothetical protein C45G9.2. - Yeast protein SMM1. - Yeast

hypothetical protein YLR401c. - Yeast hypothetical protein YLR405w. - Yeast hypothetical

protein YML080w. Although it has been proposed [2] that Rhodobacter capsulatus nifR3 is a

transcriptional regulatory protein, it is believed that these proteins constitute a family of

enzymes whose active site could include a conserved cysteine which has been used as the

20 central part of a signature pattern.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-[DNG]-[LIVM][LIVM SEQ ID NO:4]-
N-x-G-C-P-x(3)-[LIVMASQ][LIVMASQ SEQ ID NO:632]-x(5)-G-[SAC]

[1] Bairoch A., Rudd K.E. Unpublished observations (1995).[2] Foster-Hartnett D., Cullen
P.J., Gabbert K.K., Kranz R.G. Mol. Microbiol. 8:903-914(1993).

25

703. Uncharacterized protein family UPF0038 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli hypothetical protein yacE and HI0890, the corresponding Haemophilus

30 influenzae protein. - Mycobacterium tuberculosis hypothetical protein MtCY01B2.23 and

O410, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803

hypothetical protein slr0553. - Other hypothetical proteins from Aeromonas hydrophila,

Bacteroides nodosus, Neisseria gonorrhoeae, Pseudomonas putida, Thermus thermophilus

and *Xanthomonas campestris*. - Human hypothetical protein pOV-2. - Yeast hypothetical protein YDR196C. - *Caenorhabditis elegans* hypothetical protein T05G5.5. These proteins all contain, in their N-terminal extremity, an ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>). The size of these proteins range from 200 to 290 residues (with the exception of the Mycobacterial sequences which are 410 residues long). A conserved region some 50 residues away from the ATP-binding P-loop has been developed as a signature pattern.

Consensus pattern: G-x-[LI]-x-R-x(2)-L-x(4)-F-x(8)-[LIV]-x(5)-P-x-[LIV] -

[1] Rudd K.E., Bairoch A. Unpublished observations (1997).

704. Ubiquitin-conjugating enzymes active site

Ubiquitin-conjugating enzymes (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activated ubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2 which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3). In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions. A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue is conserved in the sequence of known UBC isozymes. That region has been used as a signature pattern.

Consensus pattern: [FYWLSP][FYWLSP SEQ ID NO:633]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV] [C is the active site residue]

[1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-

198(1990).[2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-

139(1991).[3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).

705. Uroporphyrinogen decarboxylase signatures

Uroporphyrinogen decarboxylase (URO-D), the fifth enzyme of the heme biosynthetic pathway, catalyzes the sequential decarboxylation of the four acetyl side chains of uroporphyrinogen to yield coproporphyrinogen [1]. URO-D deficiency is responsible for the Human genetic diseases familial porphyria cutanea tarda (fPCT) and hepatoerythropoietic porphyria (HEP). The sequence of URO-D has been well conserved throughout evolution.

The best conserved region is located in the N-terminal section; it contains a perfectly conserved hexapeptide. There are two arginine residues in this hexapeptide which could be involved in the binding, via salt bridges, to the carboxyl groups of the propionate side chains of the substrate. This region has been used as a signature pattern. A second signature pattern is based on another well conserved region which is located in the central section of the protein.

Consensus pattern: P-x-W-x-M-R-Q-A-G-R

Consensus pattern: G-F-[STAGCV][STAGCV SEQ ID NO:159]-[STAGC][STAGC SEQ ID NO:45]-x-P-[FYW]-T-[LV]-x(2)-Y-x(2)-[AE]-[GK]

[1] Garey J.R., Labbe-Bois R., Chelstowska A., Rytka J., Harrison L., Kushner J., Labbe P. Eur. J. Biochem. 205:1011-1016(1992).

706. ubiE/COQ5 methyltransferase family signatures

The following methyltransferases have been shown [1] to share regions of similarities: - Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-methoxy-1,4-benzoquinol into 2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinol and of demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis methyltransferase. - Bacillus subtilis spore germination protein C2 (gene: gercB or gerC2), a probable menaquinone biosynthesis methyltransferase. - Lactococcus lactis gerC2 homolog. - Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-specific protein A41. These are hydrophilic proteins of about 30 Kd (except for ZK652.9 which is 65Kd). They can be picked up in the database by the following patterns.

Consensus pattern: Y-D-x-M-N-x(2)-[LIVM][LIVM SEQ ID NO:4]-S-x(3)-H-x(2)-W

Consensus pattern: R-V-[LIVM][LIVM SEQ ID NO:4]-K-[PV]-G-G-x-[LIVME][LIVME SEQ ID NO:2]-x(2)-[LIVM][LIVM SEQ ID NO:4]-E-x-S

[1] Lee P.T., Hsu A.Y., Ha H.T., Clarke C.F. J. Bacteriol. 179:1748-1754(1997).

707. Uricase signature

Uricase (urate oxidase) [1] is the peroxisomal enzyme responsible for the degradation of urate into allantoin. Some species, like primates and birds, have lost the gene for uricase and

are therefore unable to degradeurate. Uricase is a protein of 300 to 400 amino acids. A highly conserved region located in the central part of the sequence has been used as a signature pattern.

Consensus pattern: [LV]-x-[LV]-[LIV]-K-[STV]-[ST]-x-[SN]-x-F-x(2)-[FY]-x(4)- [FY]-x(2)-L-x(5)-R

[1] Motojima K., Kanaya S., Goto S. J. Biol. Chem. 263:16677-16681(1988).

708. Universal stress protein family (Usp)

By a wide range of stress conditions members of the Usp family are predicted to be related to the MADS-box proteins transcript_fact and bind to DNA [2]. Number of members: 39

[1] Expression and role of the universal stress protein, UspA, of Escherichia coli during growth arrest. Nystrom T, Neidhardt FC; Mol Microbiol 1994; 11:537-544.

[2] Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. Mushegian AR, Koonin EV; Genetics 1996; 144:817-828.

709. Ubiquitin domain signature and profile

Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis. In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a Xenopus gene). In the majority of polyubiquitin precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins. Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg- Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is

mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins. There are a number of proteins which are evolutionary related to ubiquitin: -

Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. -

Mammalian protein GDX [4]. GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein

which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein

S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein

BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like

domain. - Caenorhabditis elegans protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of

a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast

DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be

distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins

RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1). BAG-1 is a

protein of 274 residues that contains a central ubiquitin-like domain. - Human spliceosome

associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in

spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. -

Human protein CKAP1/TFCB, Schizosaccharomyces pombe protein alp11 and

Caenorhabditis elegans hypothetical protein F53F4.3. These proteins contain a N-terminal

ubiquitin domain and a C-terminal CAP-Gly domain. - Schizosaccharomyces pombe

hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. -

Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human

ubiquitin-like protein SMT3C (also known as PIC1; Ubl1, Sumo-1; Gmp-1 or Sentrin). This

protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. -

SMT3-like proteins in plants and Caenorhabditis elegans. To identify ubiquitin and related

proteins, a pattern has been developed based on conserved positions in the central section of

the sequence. A profile was also developed that spans the complete length of the ubiquitin

domain.

Consensus pattern: K-x(2)-[LIVM][LIVM SEQ ID NO:4]-x-[DESAK][DESAK SEQ ID NO:634]-x(3)-[LIVM][LIVM SEQ ID NO:4]-[PA]-x(3)-Q-x-[LIVM][LIVM SEQ ID

NO:4)]- [LIVMC][LIVMC SEQ ID NO:142)]-[LIVMFY][LIVMFY SEQ ID NO:18)]-x-G-x(4)-[DE]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] Monia B.P., Ecker D.J., Croke S.T. Bio/Technology 8:209-215(1990).[3] Finley D., Varshavsky A. Trends Biochem. Sci. 10:343-347(1985).[4] Filippi M., Tribioli C., Toniolo D. Genomics 7:453-457(1990).[5] Olvera J., Wool I.G. J. Biol. Chem. 268:17967-17974(1993).[6] Kumar S., Yoshida Y., Noda M. Biochem. Biophys. Res. Commun. 195:393-399(1993).[7] Jones D., Candido E.P. J. Biol. Chem. 268:19545-19551(1993).[8] Melnick L., Sherman F. J. Mol. Biol. 233:372-388(1993).

710. VHS domain

Domain present in VPS-27, Hrs and STAM. Number of members: 27

711. Vinculin family signatures

Vinculin [1] is a eukaryotic protein that seems to be involved in the attachment of the actin-based microfilaments to the plasma membrane. Vinculin is located at the cytoplasmic side of focal contacts or adhesion plaques. In addition to actin, vinculin interacts with other structural proteins such as talin and alpha-actinins. Vinculin is a large protein of 116 Kd (about a 1000 residues). Structurally the protein consists of an acidic N-terminal domain of about 90 Kd separated from a basic C-terminal domain of about 25 Kd by a proline-rich region of about 50 residues. The central part of the N-terminal domain consists of a variable number (3 in vertebrates, 2 in *Caenorhabditis elegans*) of repeats of a 110 amino acids domain. Catenins [2] are proteins that associate with the cytoplasmic domain of a variety of cadherins. The association of catenins to cadherins produces a complex which is linked to the actin filament network, and which seems to be of primary importance for cadherins cell-adhesion properties. Three different types of catenins seem to exist: alpha, beta, and gamma. Alpha-catenins are proteins of about 100 Kd which are evolutionary related to vinculin. In terms of their structure the most significant differences are the absence, in alpha-catenin, of the repeated domain and of the proline-rich segment. Two signature patterns for this family of proteins have been developed. The first pattern is located in the N-terminal section of both vinculin and alpha-catenins and is part, in vinculin, of a domain that seems to be involved

581

with the interaction with talin. The second pattern is based on a conserved region in the N-terminal part of the repeated domain of vinculin.

Consensus pattern: [KR]-x-[LIVMF][LIVMF SEQ ID NO:2)]-x(3)-[LIVMA][LIVMA SEQ ID NO:30)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]-x(6)-R-Q-Q-E-L

5 Consensus pattern: [LIVM][LIVM SEQ ID NO:4)]-x-[QA]-A-x(2)-W-[IL]-x-[DN]-P

[1] Otto J.J. Cell Motil. Cytoskeleton 16:1-6(1990).[2] Herrenknecht K., Ozawa M., Eckerskorn C., Lottspeich F., Lenter M., Kemler R. Proc. Natl. Acad. Sci. U.S.A. 88:9156-9160(1991).

10

712. (Vitellogenin N) Lipoprotein amino terminal region

This family contains regions from: Vitellogenin, Microsomal triglyceride transfer protein and apolipoprotein B-100. These proteins are all involved in lipid transport [1]. This family contains the LV1n chain from lipovitellin, that contains two structural domains.

15 Number of members: 33

[1] The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. Anderson TA, Levitt DG, Banaszak LJ Structure 1998;6:895-909.

20 713. (VMSA) Major surface antigen from hepadnavirus

714. ssDNA binding protein (Viral DNA bp)

This protein is found in herpesviruses and is needed for replication.

25

715. (Votage CLC) Voltage gated chloride channels

30 This family of ion channels contains 10 or 12 transmembrane helices. Each protein forms a single pore. It has been shown that some members of this family form homodimers. These proteins contain two CBS domains.

[1] Schmidt-Rose T, Jentsch TJ; J Biol Chem 1997;272:20515-20521.

[2] Zhang J, George AL Jr, Griggs RC, Fouad GT, Roberts J, Kwiecinski H, Connolly AM, Ptacek LJ; Neurology 1996;47:993-998.

5

716. von Willebrand factor type A domain (vwa)

More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-sensitive calcium channel and inter-alpha-trypsin inhibitor.

10

Bork P, Rohde K;

Biochem J 1991;279:908-911.

1. RUGGERI, Z.M. and WARE, J.

15

von Willebrand factor.

FASEB J. 7 308-316 (1993).

2. COLOMBATTI, A., BONALDO, P. and DOLIANA, R.

Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins.

20

MATRIX 13 297-306 (1993).

3. PERKINS, S.J., SMITH, K.F., WILLIAMS, S.C., HARIS, P.I., CHAPMAN, D. and SIM, R.B.

25

The secondary structure of the von Willebrand factor type A domain in factor B of human complement by Fourier transform infrared spectroscopy.

Its occurrence in collagen types VI, VII, XII and XIV, the integrins and other proteins by averaged structure predictions.

J.MOL.BIOL. 238 104-119 (1994).

30

4. BORK, P. and ROHDE, K.

More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-

sensitive calcium channel and inter-alpha-trypsin inhibitor.
BIOCHEM.J. 279 908-910 (1991).

5. EDWARDS, Y.J.K. and PERKINS, S.J.

- 5 The protein fold of the von Willebrand factor type A domain is predicted to be similar to the open twisted beta-sheet flanked by alpha-helices found in human ras-p21.
FEBS LETT. 358 283-286 (1995).

- 10 6. LEE, J.O., RIEU, P., ARNAOUT, M.A. and LIDDINGTON, R.
Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18).
CELL 80 631-638 (1995).

- 15 7. QU, A. and LEAHY, D.J.
Crystal structure of the I-domain from the CD11a/CD18 (LFA-1, alpha L beta 2) integrin.
PROC.NATL.ACAD.SCI.USA 92 10277-10281 (1995).

- 20 The von Willebrand factor is a large multimeric glycoprotein found in blood plasma. Mutant forms are involved in the aetiology of bleeding disorders [1]. In von Willebrand factor, the type A domain (vWF) is the prototype for a protein superfamily. The vWF domain is found in various plasma proteins: complement factors B, C2, CR3 and CR4; the integrins (I-domains); collagen
25 types VI, VII, XII and XIV; and other extracellular proteins [2-4]. Proteins that incorporate vWF domains participate in numerous biological events (e.g., cell adhesion, migration, homing, pattern formation, and signal transduction), involving interaction with a large array of ligands [2].
Secondary structure prediction from 75 aligned vWF sequences has revealed
30 a largely alternating sequence of alpha-helices and beta-strands [3]. Fold recognition algorithms were used to score sequence compatibility with a library of known structures: the vWF domain fold was predicted to be a doubly-wound, open, twisted beta-sheet flanked by alpha-helices [5].

3D structures have been determined for the I-domains of integrins CD11b (with bound magnesium) [6] and CD11a (with bound manganese) [7]. The domain adopts a classic alpha/beta Rossmann fold and contains an unusual metal ion coordination site at its surface. It has been suggested that this site represents a general metal ion-dependent adhesion site (MIDAS) for binding protein ligands [6]. The residues constituting the MIDAS motif in the CD11b and CD11a I-domains are completely conserved, but the manner in which the metal ion is coordinated differs slightly [7].

VWFADOMAIN is a 3-element fingerprint that provides a signature for the vWF domain superfamily. The fingerprint was derived from an initial alignment of 14 sequences. Motif 1 includes the first beta-strand and 3 conserved residues involved in metal ion coordination in I-domains (Asp and 2 serines in positions 8, 10 and 12, respectively); motif 2 spans strands beta-2 and beta-2'; and motif 3 encodes beta-strand 3 and a conserved Asp (in position 7), which coordinates the metal ion [6,7]. Three iterations on OWL27.0 were required to reach convergence, at which point a true set comprising 56 sequences was identified. Numerous partial matches were also found.

717. (WD40) WD domain, G-beta repeat

The ancient regulatory-protein family of WD-repeat proteins.

Neer EJ, Schmidt CJ, Nambudripad R, Smith TF;

Nature 1994;371:297-300.

Beta-transducin (G-beta) is one of the three subunits (alpha, beta, and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane receptors [1]. The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

In higher eukaryotes G-beta exists as a small multigene family of highly

conserved proteins of about 340 amino acid residues. Structurally G-beta consists of eight tandem repeats of about 40 residues, each containing a central Trp-Asp motif (this type of repeat is sometimes called a WD-40 repeat). Such a repetitive segment has been shown [E1,2,3,4,5] to exist in a number of other proteins listed below:

- Yeast STE4, a component of the pheromone response pathway. STE4 is a G-beta like protein that associates with GPA1 (G-alpha) and STE18 (G-gamma).
- Yeast MSI1, a negative regulator of RAS-mediated cAMP synthesis. MSI1 is most probably also a G-beta protein.
- Human and chicken protein 12.3. The function of this protein is not known, but on the basis of its similarity to G-beta proteins, it may also function in signal transduction.
- *Chlamydomonas reinhardtii* gblp. This protein is most probably the homolog of vertebrate protein 12.3.
- Human LIS1, a neuronal protein involved in type-1 lissencephaly [E2].
- Mammalian coatamer beta' subunit (beta'-COP), a component of a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport.
- Yeast CDC4, essential for initiation of DNA replication and separation of the spindle pole bodies to form the poles of the mitotic spindle.
- Yeast CDC20, a protein required for two microtubule-dependent processes: nuclear movements prior to anaphase and chromosome separation.
- Yeast MAK11, essential for cell growth and for the replication of M1 double-stranded RNA.
- Yeast PRP4, a component of the U4/U6 small nuclear ribonucleoprotein with a probable role in mRNA splicing.
- Yeast PWP1, a protein of unknown function.
- Yeast SKI8, a protein essential for controlling the propagation of double-stranded RNA.
- Yeast SOF1, a protein required for ribosomal RNA processing which

associates with U3 small nucleolar RNA.

- Yeast TUP1 (also known as AER2 or SFL2 or CYC9), a protein which has been implicated in dTMP uptake, catabolite repression, mating sterility, and many other phenotypes.

5 - Yeast YCR57c, an ORF of unknown function from chromosome III.

- Yeast YCR72c, an ORF of unknown function from chromosome III.

- Slime mold coronin, an actin-binding protein.

- Slime mold AAC3, a developmentally regulated protein of unknown function.

10

- Drosophila protein Groucho (formerly known as E(spl); 'enhancer of split'), a protein involved in neurogenesis and that seems to interact with the Notch and Delta proteins.

- Drosophila TAF-II-80, a protein that is tightly associated with TFIID.

15

The number of repeats in the above proteins varies between 5 (PRP4, TUP1, and Groucho) and 8 (G-beta, STE4, MSII, AAC3, CDC4, PWP1, etc.). In G-beta and G-beta like proteins, the repeats span the entire length of the sequence, while in other proteins, they make up the N-terminal, the central or the C-terminal

20

A signature pattern can be developed from the central core of the domain (positions 9 to 23).

25

-Consensus pattern: [LIVMSTAC][LIVMSTAC SEQ ID NO:151]-
[LIVMFYWSTAGC][LIVMFYWSTAGC SEQ ID NO:635]-[LIMSTAG][LIMSTAG SEQ
ID NO:636)]-[LIVMSTAGC][LIVMSTAGC SEQ ID NO:637)]-x(2)-[DN]-
x(2)-[LIVMWSTAC][LIVMWSTAC SEQ ID NO:638)]-x-[LIVMFSTAG][LIVMFSTAG
SEQ ID NO:639)]-W-[DEN]-[LIVMFSTAGCN][LIVMFSTAGCN SEQ ID NO:640)]

30

[1] Gilman A.G.

Annu. Rev. Biochem. 56:615-649(1987).

[2] Duronio R.J., Gordon J.I., Boguski M.S.

Proteins 13:41-56(1992).

[3] van der Voorn L., Ploegh H.L.

FEBS Lett. 307:131-134(1992).

[4] Neer E.J., Schmidt C.J., Nambudripad R., Smith T.F.

5 Nature 371:297-300(1994).

[5] Smith T.F., Gaiatzes C.G., Saxena K., Neer E.J.

Biochemistry In Press(1998).

10 718. WHEP-TRS domain containing proteins

A conserved domain of 46 amino acids has been shown [1] to exist in a number of higher eukaryote aminoacyl-transfer RNA synthetases. This domain is present one to six times in the following enzymes:

15 - Mammalian multifunctional aminoacyl-tRNA synthetase. The domain is present three times in a region that separates the N-terminal glutamyl-tRNA synthetase domain from the C-terminal prolyl-tRNA synthetase domain.

- Drosophila multifunctional aminoacyl-tRNA synthetase. The domain is present six times in the intercatalytic region.

20 - Mammalian tryptophanyl-tRNA synthetase. The domain is found at the N-terminal extremity.

- Mammalian, insect, nematode and plant glycyl-tRNA synthetase. The domain is found at the N-terminal extremity [2].

25 - Mammalian histidyl-tRNA synthetase. The domain is found at the N-terminal extremity.

This domain, which is called WHEP-TRS, could contain a central alpha-helical region and may play a role in the association of tRNA-synthetases into multienzyme complexes.

30

A signature pattern based on the first 29 positions of the WHEP-Domain has been developed.

-Consensus pattern: [QY]-G-[DNEA][DNEA SEQ ID NO:641]]-x-[LIV]-[KR]-x(2)-K-x(2)-
[KRNG][KRNG SEQ ID NO:642]]-[AS]-x(4)-
[LIV]-[DENK][DENK SEQ ID NO:643]]-x(2)-[IV]-x(2)-L-x(3)-K

- 5 [1] Cerini C., Kerjan P., Astier M., Gratecos D., Mirande M., Semeriva M.
EMBO J. 10:4267-4277(1991).
[2] Nada S., Chang P.K., Dignam J.D.
J. Biol. Chem. 268:7660-7667(1993).

10

719. (Worm family 8) Putative membrane protein
Analysis of protein domain families in *Caenorhabditis elegans*.
Sonnhammer EL, Durbin R;
Genomics 1997;46:200-216.

- 15 This family called family 8 in [1], may be a transmembrane protein
The specific function of this protein is unknown.

720. Xylose isomerase

- 20 Xylose isomerase (EC 5.3.1.5) [1] is an enzyme found in microorganisms which
catalyzes the interconversion of D-xylose to D-xylulose. It can also isomerize
D-ribose to D-ribulose and D-glucose to D-fructose. Xylose isomerase seems to
require magnesium for its activity, while cobalt is necessary to stabilize the
tetrameric structure of the enzyme. A number of residues are conserved in all
25 known xylose isomerases.

Xylose isomerase also exists in plants [2] where it is homodimeric and is
manganese-dependent.

- 30 Two signatures patterns for xylose isomerase have been developed. The first one is
derived from a stretch of five conserved amino acids that includes a glutamic
acid residue known to be one of the four residues involved in the binding of
the magnesium ion [3]; this pattern also includes a lysine residue which is

involved in the catalytic activity. The second pattern is derived from a conserved region in the N-terminal section of the enzyme that include an histidine residue which has been shown [4] to be involved in the catalytic mechanism of the enzyme.

5

-Consensus pattern: [LI]-E-P-K-P-x(2)-P

[E is a magnesium ligand]

[K is an active site residue]

-Consensus pattern: [FL]-H-D-x-D-[LIV]-x-[PD]-x-[GDE]

10

[H is an active site residue]

[1] Dauter Z., Dauter M., Hemker J., Witzel H., Wilson K.S.

FEBS Lett. 247:1-8(1989).

[2] Kristo P.A., Saarelainen R., Fagerstrom R., Aho S., Korhola M.

15

Eur. J. Biochem. 237:240-246(1996).

[3] Henrick K., Collyer C.A., Blow D.M.

J. Mol. Biol. 208:129-157(1989).

[4] Vangrysperre W., Ampe C., Kersters-Hilderson H., Tempst P.

Biochem. J. 263:195-199(1989).

20

721. XPG protein signatures. Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2]. XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets: - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3' incision in human DNA nucleotide excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. In addition to the proteins listed in the above groups, this family also includes: - Fission yeast

25

30

exo1, a 5'→3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as exo1. - Yeast DIN7. Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset. Two signature patterns have been developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide

Consensus pattern: [VI]-[KRE]-P-x-[FYIL][FYIL SEQ ID NO:644]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K-

Consensus pattern: [GS]-[LIVM][LIVM SEQ ID NO:4]-[PER]-[FYS]-[LIVM][LIVM SEQ ID NO:4]-x-A-P-x-E-A-[DE]-[PAS]-[QS]-[CLM]-

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).[2] Scherly D., Nouspikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).[3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).[4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).[5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).[6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

722. Xanthine/uracil permeases family

The following transport proteins which are involved in the uptake of xanthine

or uracil are evolutionary related [1]:

- Uric uric acid-xanthine permease (gene uapA) from *Aspergillus nidulans*.
- Purine permease (gene uapC) from *Aspergillus nidulans*.
- 5 - Xanthine permease from *Bacillus subtilis* (gene pbuX).
- Uracil permease from *Escherichia coli* (gene uraA) [2] and *Bacillus* (gene pyrP).
- Hypothetical protein ycdG from *Escherichia coli*.
- Hypothetical protein ygfO from *Escherichia coli*.
- 10 - Hypothetical protein ygfU from *Escherichia coli*.
- Hypothetical protein yicE from *Escherichia coli*.
- Hypothetical protein yunJ from *Bacillus subtilis*.
- Hypothetical protein yunK from *Bacillus subtilis*.

15 They are proteins of from 430 to 595 residues that seem to contain 12 transmembrane domains.

The best conserved region which corresponds with what seems to be the tenth transmembrane domain has been selected as a signature pattern.

20 -Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-P-x-[PASIF][PASIF SEQ ID NO:645]-V-[LIVM][LIVM SEQ ID NO:4]-G-G-x(4)-[LIVM][LIVM SEQ ID NO:4]-[FY]-[GSA]-x-[LIVM][LIVM SEQ ID NO:4]-x(3)-G

[1] Diallinas G., Gorfinkiel L., Arst G., Cecchetto G., Scazzocchio C.
J. Biol. Chem. 270:8610-8622(1995).

25 [2] Andersen P.S., Frees D., Fast R., Mygind B.
J. Bacteriol. 177:2008-2013(1995).

723. Hypothetical yabO/yceC/sfhB family

30 The following proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] have been shown to share regions of similarities:

- *Escherichia coli* and *Haemophilus influenzae* ribosomal large subunit

pseudouridine synthase A (gene rluA). It is responsible for synthesis of pseudouridine from uracil-746 IN 23S rRNA.

- Escherichia coli and Haemophilus influenzae ribosomal large subunit

pseudouridine synthase C (gene rluC). It is responsible for synthesis of

5 pseudouridine from uracil at positions 955, 2504 and 2580 in 23S rRNA.

- Escherichia coli protein and homologs in other bacteria large subunit

pseudouridine synthase D (gene rluD).

- Yeast DRAP deaminase (gene RIB2).

- Escherichia coli hypothetical protein yqcB and HI1435, the corresponding

10 Haemophilus influenzae protein.

- Haemophilus influenzae hypothetical protein HI0042.

- Aquifex aeolicus hypothetical protein AQ_1758.

- Bacillus subtilis hypothetical protein yhcT.

- Bacillus subtilis hypothetical protein yjbO.

15 - Bacillus subtilis hypothetical protein ylyB.

- Helicobacter pylori hypothetical protein HP0347.

- Helicobacter pylori hypothetical protein HP0745.

- Helicobacter pylori hypothetical protein HP0956.

- Mycoplasma genitalium hypothetical protein MG209.

20 - Mycoplasma genitalium hypothetical protein MG370.

- Synechocystis strain PCC 6803 hypothetical protein slr1592.

- Synechocystis strain PCC 6803 hypothetical protein slr1629.

- Yeast hypothetical protein YDL036c.

- Yeast hypothetical protein YGR169c.

25 - Fission yeast hypothetical protein SpAC18B11.02c.

- Caenorhabditis elegans hypothetical protein K07E8.7.

These are proteins of from 21 to 50 Kd which contain a number of conserved regions in their central section. They can be picked up in the database by the
30 following highly conserved pattern.

-Consensus pattern: [LIVCA][LIVCA SEQ ID NO:646)]-[NHYT][NHYT SEQ ID NO:647)]-R-[LI]-D-x(2)-T-[STA]-G-[LIVAGC][LIVAGC SEQ ID NO:648)]-

[LIVMF][LIVMF SEQ ID NO:2)](2)-[LIVMFGC][LIVMFGC SEQ ID NO:649)]-
[SGTACV][SGTACV SEQ ID NO:650)]

[1] Conrad J., Sun D., Englund N., Ofengand J.

5 J. Biol. Chem. 273:18562-18566(1998).

In addition, the following bacterial proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] also have been shown to share regions of similarities:

10

- Escherichia coli and Haemophilus influenzae 16S pseudouridylate 516 synthase (EC 4.2.1.70) (gene: rsuA). This enzyme is responsible for the formation of pseudouridine from uracil-516 in 16S ribosomal RNA.

15

- Escherichia coli hypothetical protein yciL and HI1199, the corresponding Haemophilus influenzae protein.

- Escherichia coli hypothetical protein yjbC.

- Escherichia coli hypothetical protein ymfC and HI0694, the corresponding Haemophilus influenzae protein.

- Aquifex aeolicus hypothetical protein AQ_554.

20

- Aquifex aeolicus hypothetical protein AQ_1464.

- Bacillus subtilis hypothetical protein ypuL.

- Bacillus subtilis hypothetical protein ytzF.

- Borrelia burgdorferi hypothetical protein BB0129.

- Helicobacter pylori hypothetical protein HP1459.

25

- Synechocystis strain PCC 6803 hypothetical protein slr0361.

- Synechocystis strain PCC 6803 hypothetical protein slr0612.

These are proteins of from 25 to 40 Kd which contain a number of conserved regions in their central section. They can be picked up in the database by the
30 following highly conserved pattern.

-Consensus pattern: G-R-L-D-x(2)-[STA]-x-G-[LIVFA][LIVFA SEQ ID NO:129)]-
[LIVMF][LIVMF SEQ ID NO:2)](3)-[ST]-[DNST][DNST SEQ ID NO:265)]

[1] Wrzesinski J., Bakin A., Nurse K., Lane B.G., Ofengand J.
Biochemistry 34:8904-8913(1995).

5

724. Zinc finger present in dystrophin, CBP/p300
ZZ in dystrophin binds calmodulin
Putative zinc finger; binding not yet shown.

10

725. Zinc carboxypeptidase

There are a number of different types of zinc-dependent carboxypeptidases (EC 3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally related. The enzymes that belong to this family are listed below.

15

- Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that can removes all C-terminal amino acids with the exception of Arg, Lys and Pro.

- Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a specificity similar to that of carboxypeptidase A1, but with a preference
20 for bulkier C-terminal residues.

- Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but that preferentially removes C-terminal Arg and Lys.

- Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase), a plasma enzyme which protects the body from potent vasoactive and
25 inflammatory peptides containing C-terminal Arg or Lys (such as kinins or anaphylatoxins) which are released into the circulation.

- Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or carboxypeptidase E), an enzyme located in secretory granules of pancreatic islets, adrenal gland, pituitary and brain. This enzyme removes residual C-
30 terminal Arg or Lys remaining after initial endoprotease cleavage during prohormone processing.

- Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific enzyme.

It is ideally situated to act on peptide hormones at local tissue sites where it could control their activity before or after interaction with specific plasma membrane receptors.

- Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity to carboxypeptidase A, but found in the secretory granules of mast cells.
- *Streptomyces griseus* carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which combines the specificities of mammalian carboxypeptidases A and B.
- *Thermoactinomyces vulgaris* carboxypeptidase T (EC 3.4.17.18) (CPT) [4], which also combines the specificities of carboxypeptidases A and B.
- AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems to regulate transcription by cleavage of other transcriptional proteins.
- Yeast hypothetical protein YHR132c.

All of these enzymes bind an atom of zinc. Three conserved residues are implicated in the binding of the zinc atom: two histidines and a glutamic acid. Two signature patterns which contain these three zinc-ligands have been derived.

-Consensus pattern: [PK]-x-[LIVMFY][LIVMFY SEQ ID NO:18]-x-[LIVMFY][LIVMFY SEQ ID NO:18]-x(4)-H-[STAG][STAG SEQ ID NO:20]-x-E-x-[LIVM][LIVM SEQ ID NO:4]-[STAG][STAG SEQ ID NO:20]-x(6)-[LIVMFYTA][LIVMFYTA SEQ ID NO:651]]
[H and E are zinc ligands]

-Consensus pattern: H-[STAG][STAG SEQ ID NO:20]-x(3)-[LIVME][LIVME SEQ ID NO:652]-x(2)-[LIVMFYW][LIVMFYW SEQ ID NO:26]-P-[FYW]
[H is a zinc ligand]

[1] Tan F., Chan S.J., Steiner D.F., Schilling J.W., Skidgel R.A.
J. Biol. Chem. 264:13165-13170(1989).

[2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K.F.,
Serafin W.E.
J. Biol. Chem. 264:20094-20099(1989).

[3] Narahashi Y.
J. Biochem. 107:879-886(1990).

[4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I., Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P., Galperina O.V., Matz M.V., Stepanov V.M.
Eur. J. Biochem. 208:281-288(1992).

5 [5] He G.-P., Muise A., Li A.W., Ro H.-S.
Nature 378:92-96(1995).

[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L., Duez C.M., Englebert S., Joris B., Weber G., Ghuyssen J.-M.
Biochem. J. 292:563-570(1993).

10 [7] Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).

726. Zinc finger, C2H2 type

15 The C2H2 zinc finger is the classical zinc finger domain.

The two conserved cysteines and histidines co-ordinate a zinc ion. The following pattern describes the zinc finger.

#-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C]

Where X can be any amino acid, and numbers in brackets

20 indicate the number of residues. The positions marked # are those that are important for the stable fold of the zinc finger. The final position can be either his or cys.

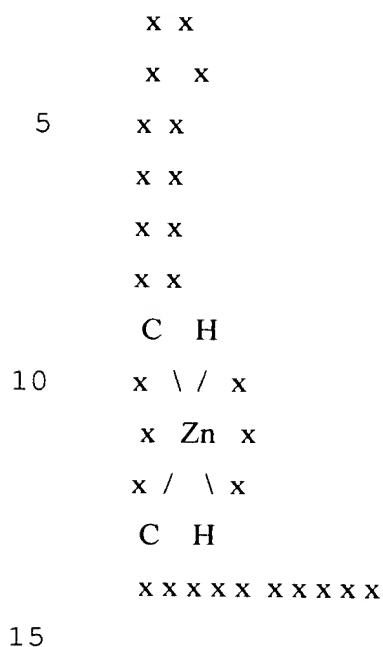
The C2H2 zinc finger is composed of two short beta strands followed by an alpha helix. The amino terminal part of the

25 helix binds the major groove in DNA binding zinc fingers.

'Zinc finger' domains [1-5] are nucleic acid-binding protein structures first identified in the *Xenopus* transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger

30 domain is composed of 25 to 30 amino-acid residues. There are two cysteine or histidine residues at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with about five nucleotides. A schematic representation of a

zinc finger domain is shown below:



Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Some of the proteins known to include C2H2-type zinc fingers are listed below. The number of zinc finger regions found in each of these proteins are indicated between brackets; a '+' symbol indicates that only partial sequence data is available and that additional finger domains may be present.

- *Saccharomyces cerevisiae*: ACE2 (3), ADR1 (2), AZF1 (4), FZF1 (5), MIG1 (2), MSN2 (2), MSN4 (2), RGM1 (2), RIM1 (3), RME1 (3), SFP1 (2), SSL1 (1), STP1 (3), SWI5 (3), VAC1 (1) and ZMS1 (2).
- *Emericella nidulans*: brlA (2), creA (2).
- *Drosophila*: AEF-1 (4), Cf2 (7), ci-D (5), Disconnected (2), Escargot (5), Glass (5), Hunchback (6), Kruppel (5), Kruppel-H (4+), Odd-skipped (4),

Odd-paired (4), Pep (3), Snail (5), Spalt-major (7), Serependity locus beta (6), delta (7), h-1 (8), Suppressor of hairy wing su(Hw) (12), Suppressor of variegation suvar(3)7 (5), Teashirt (3) and Tramtrack (2).

- Xenopus: transcription factor TFIIIA (9), p43 from RNP particle (9), Xfin (37 !!), Xsna (5), gastrula XlclGF5.1 to XlclGF71.1 (from 4+ to 11+), Oocyte XlclOF2 to XlclOF22 (from 7 to 12).

- Mammalian: basoonuclin (6), BCL-6/LAZ-3 (6), erythroid krueppel-like transcription factor (3), transcription factors Sp1 (3), Sp2 (3), Sp3 (3) and Sp(4) 3, transcriptional repressor YY1 (4), Wilms' tumor protein (4),

EGR1/Krox24 (3), EGR2/Krox20 (3), EGR3/Pilot (3), EGR4/AT133 (4), Evi-1 (10), GLI1 (5), GLI2 (4+), GLI3 (3+), HIV-EP1/ZNF40 (4), HIV-EP2 (2), KR1 (9+), KR2 (9), KR3 (15+), KR4 (14+), KR5 (11+), HF.12 (6+), REX-1 (4), Zfx (13), Zfy (13), Zfp-35 (18), ZNF7 (15), ZNF8 (7), ZNF35 (10), ZNF42/MZF-1 (13), ZNF43 (22), ZNF46/Kup (2), ZNF76 (7), ZNF91 (36), ZNF133 (3).

In addition to the conserved zinc ligand residues it has been shown [6] that a number of other positions are also important for the structural integrity of the C2H2 zinc fingers. The best conserved position is found four residues after the second cysteine; it is generally an aromatic or aliphatic residue.

-Consensus pattern: C-x(2,4)-C-x(3)-[LIVMFYWC][LIVMFYWC SEQ ID NO:86]-x(8)-H-x(3,5)-H

[The two C's and two H's are zinc ligands]

[1] Klug A., Rhodes D.

Trends Biochem. Sci. 12:464-469(1987).

[2] Evans R.M., Hollenberg S.M.

Cell 52:1-3(1988).

[3] Payre F., Vincent A.

FEBS Lett. 234:245-250(1988).

[4] Miller J., McLachlan A.D., Klug A.

EMBO J. 4:1609-1614(1985).

[5] Berg J.M.

Proc. Natl. Acad. Sci. U.S.A. 85:99-102(1988).

[6] Rosenfeld R., Margalit H.

J. Biomol. Struct. Dyn. 11:557-570(1993).

5

727. Zinc finger, C3HC4 type (RING finger)

A number of eukaryotic and viral proteins contain a conserved cysteine-rich domain of 40 to 60 residues (called C3HC4 zinc-finger or 'RING' finger) [1] that binds two atoms of zinc, and is probably involved in mediating protein-protein interactions. The 3D structure of the zinc ligation system is unique to the RING domain and is referred to as the "cross-brace" motif. The spacing of the cysteines in such a domain is C-x(2)-C-x(9 to 39)-C-x(1 to 3)-H-x(2 to 3)-C-x(2)-C-x(4 to 48)-C-x(2)-C.

10

15

Proteins currently known to include the C3HC4 domain are listed below (references are only provided for recently determined sequences).

- Mammalian V(D)J recombination activating protein (gene RAG1). RAG1 activates the rearrangement of immunoglobulin and T-cell receptor genes.

20

- Mouse rpt-1. Rpt-1 is a trans-acting factor that regulates gene expression directed by the promoter region of the interleukin-2 receptor alpha chain or the LTR promoter region of HIV-1.

- Human rfp. Rfp is a developmentally regulated protein that may function in male germ cell development. Recombination of the N-terminal section of rfp with a protein tyrosine kinase produces the ret transforming protein.

25

- Human 52 Kd Ro/SS-A protein. A protein of unknown function from the Ro/SS-A ribonucleoprotein complex. Sera from patients with systemic lupus erythematosus or primary Sjogren's syndrome often contain antibodies that react with the Ro proteins.

30

- Human histocompatibility locus protein RING1.

- Human PML, a probable transcription factor. Chromosomal translocation of PML with retinoic receptor alpha creates a fusion protein which is the cause of acute promyelocytic leukemia (APL).

- Mammalian breast cancer type 1 susceptibility protein (BRCA1) [E1].
- Mammalian cbl proto-oncogene.
- Mammalian bmi-1 proto-oncogene.
- Vertebrate CDK-activating kinase (CAK) assembly factor MAT1, a protein that
5 stabilizes the complex between the CDK7 kinase and cyclin H (MAT1 stands
for 'Menage A Trois').
- Mammalian mel-18 protein. Mel-18 which is expressed in a variety of tumor
cells is a transcriptional repressor that recognizes and bind a specific
DNA sequence.
- 10 - Mammalian peroxisome assembly factor-1 (PAF-1) (PMP35), which is somewhat
involved in the biogenesis of peroxisomes. In humans, defects in PAF-1 are
responsible for a form of Zellweger syndrome, an autosomal recessive
disorder associated with peroxisomal deficiencies.
- Human MAT1 protein, which interacts with the CDK7-cyclin H complex.
- 15 - Human RING1 protein.
- Xenopus XNF7 protein, a probable transcription factor.
- Trypanosoma protein ESAG-8 (T-LR), which may be involved in the
postranscriptional regulation of genes in VSG expression sites or may
interact with adenylate cyclase to regulate its activity.
- 20 - Drosophila proteins Posterior Sex Combs (Psc) and Suppressor two of zeste
(Su(z)2). The two proteins belong to the Polycomb group of genes needed to
maintain the segment-specific repression of homeotic selector genes.
- Drosophila protein male-specific msl-2, a DNA-binding protein which is
involved in X chromosome dosage compensation (the elevation of
transcription of the male single X chromosome).
- 25 - Arabidopsis thaliana protein COP1 which is involved in the regulation of
photomorphogenesis.
- Fungal DNA repair proteins RAD5, RAD16, RAD18 and rad8.
- Herpesviruses trans-acting transcriptional protein ICP0/IE110. This protein
30 which has been characterized in many different herpesviruses is a trans-
activator and/or -repressor of the expression of many viral and cellular
promoters.
- Baculoviruses protein CG30.

601

- Baculoviruses major immediate early protein (PE-38).
- Baculoviruses immediate-early regulatory protein IE-N/IE-2.
- Caenorhabditis elegans hypothetical proteins F54G8.4, R05D3.4 and T02C1.1.
- Yeast hypothetical proteins YER116c and YKR017c.

5

The central region of the domain was selected as a signature pattern for the C3HC4 finger.

10

-Consensus pattern: C-x-H-x-[LIVMFY][LIVMFY SEQ ID NO:18]-C-x(2)-C-[LIVMYA][LIVMYA SEQ ID NO:609]

[1] Borden K.L.B., Freemont P.S.
Curr. Opin. Struct. Biol. 6:395-401(1996).

15

728. Zinc finger C-x8-C-x5-C-x3-H type (and similar).

20

729. Zinc finger, CCHC class

A family of CCHC zinc fingers, mostly from retroviral gag proteins (nucleocapsid). Prototype structure is from HIV. Also contains members involved in eukaryotic gene regulation, such as C. elegans GLH-1.

25

Structure is an 18-residue zinc finger; no examples of indels in the alignment.

30

730. Zn-finger in Ran binding protein and others.

731. AN1-like Zinc finger

Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like protein in *Xenopus laevis*. The following pattern describes the zinc finger. C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-X2-H-X5-H-X-C Where X can be any amino acid, and numbers in brackets indicate the number of residues.

5

[1] Linnen JM, Bailey CP, Weeks DL; Gene 1993;128:181-188.

732. 14-3-3 proteins

10 Structure of a 14-3-3 protein and implications for coordination of multiple signalling pathways.

Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, Gamblin SJ; Nature 1995;376:188-191.

Crystal structure of the zeta isoform of the 14-3-3 protein.

15 Liu D, Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington R; Nature 1995;376:191-194.

Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine.

20 Muslin AJ, Tanner JW, Allen PM, Shaw AS; Cell 1996;84:889-897.

The 14-3-3 protein binds its target proteins with a common site located towards the C-terminus.

25 Ichimura T, Ito M, Itagaki C, Takahashi M, Horigome T, Omata S, Ohno S, Isobe T
FEBS Lett 1997;413:273-276.

Molecular evolution of the 14-3-3 protein family.

30 Wang W, Shakes DC
J Mol Evol 1996;43:384-398.

Function of 14-3-3 proteins.

Jin DY, Lyu MS, Kozak CA, Jeang KT

Nature 1996;382:308-308.

5 The 14-3-3 proteins [1,2,3] are a family of closely related acidic homodimeric proteins of about 30 Kd which were first identified as being very abundant in mammalian brain tissues and located preferentially in neurons. The 14-3-3 proteins seem to have multiple biological activities and play a key role in signal transduction pathways and the cell cycle. They interact with kinases such as PKC or Raf-1; they seem to also function as protein-kinase dependent activators of tyrosine and tryptophan hydroxylases and in plants they are
10 associated with a complex that binds to the G-box promoter elements.

The 14-3-3 family of proteins are ubiquitously found in all eukaryotic species studied and have been sequenced in fungi (yeast BMH1 and BMH2, fission yeast rad24 and rad25), plants, Drosophila, and vertebrates. The sequences of the
15 14-3-3 proteins are extremely well conserved. Two highly conserved regions have been selected as signature patterns: the first is a peptide of 11 residues located in the N-terminal section; the second, a 20 amino acid region located in the C-terminal section.

20 -Consensus pattern: R-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA]
-Consensus pattern: Y-K-[DE]-S-T-L-I-[IM]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-[TAN]-[SAD]

[1] Aitken A.
25 Trends Biochem. Sci. 20:95-97(1995).
[2] Morrison D.
Science 266:56-57(1994).
[3] Xiao B., Smerdon S.J., Jones D.H., Dodson G.G., Soneji Y., Aitken A.,
Gamblin S.J.
30 Nature 376:188-191(1995).

733. D-isomer specific 2-hydroxyacid dehydrogenases (2 Hacid DH)

This Pfam covers the Formate dehydrogenase, D-glycerate dehydrogenase and D-lactate dehydrogenase families in SCOP. A number of NAD-dependent 2-hydroxyacid dehydrogenases which seem to be specific for the D-isomer of their substrate have been shown [1,2,3,4] to be functionally and structurally related. These enzymes are listed below.

- D-lactate dehydrogenase (EC 1.1.1.28), a bacterial enzyme which catalyzes the reduction of D-lactate to pyruvate.
- D-glycerate dehydrogenase (EC 1.1.1.29) (NADH-dependent hydroxypyruvate reductase), a plant leaf peroxisomal enzyme that catalyzes the reduction of hydroxypyruvate to glycerate. This reaction is part of the glycolate pathway of photorespiration.
- D-glycerate dehydrogenase from the bacteria *Hyphomicrobium methylovorum* and *Methylobacterium extorquens*.
- 3-phosphoglycerate dehydrogenase (EC 1.1.1.95), a bacterial enzyme that catalyzes the oxidation of D-3-phosphoglycerate to 3-phosphohydroxypyruvate. This reaction is the first committed step in the 'phosphorylated' pathway of serine biosynthesis.
- Erythronate-4-phosphate dehydrogenase (EC 1.1.1.-) (gene *pdxB*), a bacterial enzyme involved in the biosynthesis of pyridoxine (vitamin B6).
- D-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (D-hicDH), a bacterial enzyme that catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and D-2-hydroxy-carboxylic acids.
- Formate dehydrogenase (EC 1.2.1.2) (FDH) from the bacteria *Pseudomonas* sp. 101 and various fungi [5].
- Vancomycin resistance protein *vanH* from *Enterococcus faecium*; this protein is a D-specific alpha-keto acid dehydrogenase involved in the formation of a peptidoglycan which does not terminate by D-alanine thus preventing vancomycin binding.
- *Escherichia coli* hypothetical protein *ycdW*.
- *Escherichia coli* hypothetical protein *yiaE*.
- *Haemophilus influenzae* hypothetical protein HI1556.
- Yeast hypothetical protein YER081w.
- Yeast hypothetical protein YIL074w.

All these enzymes have similar enzymatic activities and are structurally related. Three of the most conserved regions of these proteins have been selected to develop patterns. The

first pattern is based on a glycine-rich region located in the central section of these enzymes; this region probably corresponds to the NAD-binding domain. The two other patterns contain a number of conserved charged residues, some of which may play a role in the catalytic mechanism.

5

-Consensus pattern: [LIVMA][LIVMA SEQ ID NO:30)]-[AG]-[IVT]-[LIVMFY][LIVMFY SEQ ID NO:18)]-[AG]-x-G-[NHKRQGSAC][NHKRQGSAC SEQ ID NO:653)]-[LIV]-G-x(13,14)-[LIVMT][LIVMT SEQ ID NO:654)]-x(2)-[FYwCTH][FYwCTH SEQ ID NO:655)]-[DNSTK][DNSTK SEQ ID NO:656)]

10

-Consensus pattern: [LIVMFYWA][LIVMFYWA SEQ ID NO:41)]-[LIVFYWC][LIVFYWC SEQ ID NO:657)]-x(2)-[SAC]-[DNQHR][DNQHR SEQ ID NO:658)]-[IVFA][IVFA SEQ ID NO:659)]-[LIVF][LIVF SEQ ID NO:127)]-x-[LIVF][LIVF SEQ ID NO:127)]-[HNI]-x-P-x(4)-[STN]-x(2)-[LIVMF][LIVMF SEQ ID NO:2)]-x-[GSDN][GSDN SEQ ID NO:660)]

15

-Consensus pattern: [LMFATC][LMFATC SEQ ID NO:661)]-[KPQ]-x-[GSTDN][GSTDN SEQ ID NO:662)]-x-[LIVMFYWR][LIVMFYWR SEQ ID NO:85)]-[LIVMFYW][LIVMFYW SEQ ID NO:26)](2)-N-x-[STAGC][STAGC SEQ ID NO:45)]-R-[GP]-x-[LIVH][LIVH SEQ ID NO:663)]-[LIVMC][LIVMC SEQ ID NO:142)]-[DNV]

20

[1] Grant G.A. Biochem. Biophys. Res. Commun. 165:1371-1374(1989).

[2] Kochhar S., Hunziker P., Leong-Morgenthaler P.M., Hottinger H. Biochem. Biophys. Res. Commun. 184:60-66(1992).

[3] Ohta T., Taguchi H. J. Biol. Chem. 266:12588-12594(1991).

[4] Goldberg J.D., Yoshida T., Brick P. J. Mol. Biol. 236:1123-1140(1994).

25

[5] Popov V.O., Lamzin V.S. Biochem. J. 301:625-643(1994).

734. 2-oxo acid dehydrogenases acyltransferase (catalytic domain)

Refined crystal structure of the catalytic domain of dihydrolipoyl

30

transacetylase (E2P) from azotobacter vineelandii at 2.6 angstroms resolution.

Mattevi A, Obmolova G, Kalk KH, Westphal AH, De Kok A, Hol WG;

J Mol Biol 1993;230:1183-1199.

These proteins contain one to three copies of a lipoyl binding domain followed by the catalytic domain.

- 5 735. 3-beta hydroxysteroid dehydrogenase/isomerase family
Structure and tissue-specific expression of 3
beta-hydroxysteroid dehydrogenase/5-ene-4-ene isomerase
genes in human and rat classical and peripheral
steroidogenic tissues.
- 10 Labrie F, Simard J, Luu-The V, Pelletier G, Belanger A,
Lachance Y, Zhao HF, Labrie C, Breton N, de Launoit Y, et al
J Steroid Biochem Mol Biol 1992;41:421-435.
The enzyme 3 beta-hydroxysteroid dehydrogenase/5-ene-4-ene
isomerase (3 beta-HSD) catalyzes the oxidation and isomerization
15 of 5-ene-3 beta-hydroxypregnene and 5-ene-hydroxyandrostene
steroid precursors into the corresponding 4-ene-ketosteroids necessary
for the formation of all classes of steroid hormones.
- 20 736. 3-hydroxyacyl-CoA dehydrogenase
This family also includes lambda crystallin.
Structure of L-3-hydroxyacyl-coenzyme A dehydrogenase:
preliminary chain tracing at 2.8-A resolution.
Birktoft JJ, Holden HM, Hamlin R, Xuong NH, Banaszak LJ;
25 Proc Natl Acad Sci U S A 1987;84:8262-8266.

3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to 3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems,
30 one located in mitochondria and the other in peroxisomes. In peroxisomes 3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and 3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-terminal domain bears the hydratase/isomerase activities and the C-terminal

domain the dehydrogenase activity. There are two mitochondrial enzymes: one which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.

- 5 In *Escherichia coli* (gene *fadB*) and *Pseudomonas fragi* (gene *faoA*) HCDH is part of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2].

The other proteins structurally related to HCDH are:

- 10 - Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3].
 - Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit).

15 There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. A signature pattern has been derived from this central region.

20 -Consensus pattern: [DNE]-x(2)-[GA]-F-[LIVMFY][LIVMFY SEQ ID NO:18)]-x-[NT]-R-x(3)-[PA]-[LIVMFY][LIVMFY SEQ ID NO:18)](2)-x(5)-[LIVMFYCT][LIVMFYCT SEQ ID NO:447)]-[LIVMFY][LIVMFY SEQ ID NO:18)]-x(2)-[GV]

25 [1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J.
 Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).

[2] Nakahigashi K., Inokuchi H.
 Nucleic Acids Res. 18:4937-4937(1990).

30 [3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A.,
 Tabaqchali S.
 FEMS Microbiol. Lett. 124:61-67(1994).

[4] Mulders J.W.M., Hendriks W., Blankesteyn W.M., Bloemendal H.,

de Jong W.W.

J. Biol. Chem. 263:15462-15466(1988).

5 737. 60s Acidic ribosomal protein

Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects.

Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Rodriguez Gabriel MA, Guarinos E, Ballesta JP;

10 Biochem Cell Biol 1995;73:959-968.

This family includes archaeobacterial L12, eukaryotic P0, P1 and P2.

738. 6-phosphogluconate dehydrogenases

15 6-phosphogluconate dehydrogenase (EC 1.1.1.44) (6PGD) catalyzes the third step in the hexose monophosphate shunt, the decarboxylating reduction of 6-phosphogluconate in to ribulose 5-phosphate.

Prokaryotic and eukaryotic 6PGD are proteins of about 470 amino acids whose
20 sequence are highly conserved [1]. A region which has been shown [2], from studies of the sheep 6PGD tertiary structure, to be involved in the binding of 6-phosphogluconate has been selected as a signature pattern.

-Consensus pattern: [LIVM][LIVM SEQ ID NO:4]-x-D-x(2)-[GA]-[NQS]-K-G-T-G-x-W

25

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J.
Mol. Microbiol. 5:1081-1089(1991).

[2] Adams M.J., Archibald I.G., Bugg C.E., Carne A., Gover S., Helliwell J.R., Pickersgill R.W., White S.W.

30 EMBO J. 2:1009-1014(1983).

739. (7tm 1) G-protein coupled receptors [1 to 4,E1,E2] (also called R7G) are an extensive

group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. The receptors that are currently known to belong to this family are listed below.

5

- 5-hydroxytryptamine (serotonin) 1A to 1F, 2A to 2C, 4, 5A, 5B, 6 and 7 [5].
- Acetylcholine, muscarinic-type, M1 to M5.
- Adenosine A1, A2A, A2B and A3 [6].
- Adrenergic alpha-1A to -1C; alpha-2A to -2D; beta-1 to -3 [7].

10

- Angiotensin II types I and II.
- Bombesin subtypes 3 and 4.
- Bradykinin B1 and B2.
- c3a and C5a anaphylatoxin.
- Cannabinoid CB1 and CB2.

15

- Chemokines C-C CC-CKR-1 to CC-CKR-8.
- Chemokines C-X-C CXC-CKR-1 to CXC-CKR-4.
- Cholecystokinin-A and cholecystokinin-B/gastrin.
- Dopamine D1 to D5 [8].
- Endothelin ET-a and ET-b [9].

20

- fMet-Leu-Phe (fMLP) (N-formyl peptide).
- Follicle stimulating hormone (FSH-R) [10].
- Galanin.
- Gastrin-releasing peptide (GRP-R).
- Gonadotropin-releasing hormone (GNRH-R).

25

- Histamine H1 and H2 (gastric receptor I).
- Lutropin-choriogonadotropic hormone (LSH-R) [10].
- Melanocortin MC1R to MC5R.
- Melatonin.

30

- Neuromedin B (NMB-R).
- Neuromedin K (NK-3R).
- Neuropeptide Y types 1 to 6.
- Neurotensin (NT-R).
- Octopamine (tyramine), from insects.

- Odorants [11].
- Opioids delta-, kappa- and mu-types [12].
- Oxytocin (OT-R).
- Platelet activating factor (PAF-R).
- 5 - Prostacyclin.
- Prostaglandin D2.
- Prostaglandin E2, EP1 to EP4 subtypes.
- Prostaglandin F2.
- Purinoreceptors (ATP) [13].
- 10 - Somatostatin types 1 to 5.
- Substance-K (NK-2R).
- Substance-P (NK-1R).
- Thrombin.
- Thromboxane A2.
- 15 - Thyrotropin (TSH-R) [10].
- Thyrotropin releasing factor (TRH-R).
- Vasopressin V1a, V1b and V2.
- Visual pigments (opsins and rhodopsin) [14].
- Proto-oncogene mas.
- 20 - A number of orphan receptors (whose ligand is not known) from mammals and birds.
- *Caenorhabditis elegans* putative receptors C06G4.5, C38C10.1, C43C3.2, T27D1.3 and ZC84.4.
- Three putative receptors encoded in the genome of cytomegalovirus: US27,
- 25 US28, and UL33.
- ECRF3, a putative receptor encoded in the genome of herpesvirus saimiri.

The structure of all these receptors is thought to be identical. They have seven hydrophobic regions, each of which most probably spans the membrane.

- 30 The N-terminus is located on the extracellular side of the membrane and is often glycosylated, while the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three intracellular loops to link the seven transmembrane regions. Most, but not all of these

receptors, lack a signal peptide. The most conserved parts of these proteins are the transmembrane regions and the first two cytoplasmic loops. A conserved acidic-Arg-aromatic triplet is present in the N-terminal extremity of the second cytoplasmic loop [15] and could be implicated in the interaction with G proteins.

To detect this widespread family of proteins, a pattern that contains the conserved triplet and that also spans the major part of the third transmembrane helix has been developed.

-Consensus pattern: {GSTALIVMFYWC}[GSTALIVMFYWC SEQ ID NO:664)]-[GSTANCPDE][GSTANCPDE SEQ ID NO:665)]-{EDPKRH}[EDPKRH SEQ ID NO:666)]-x(2)-[LIVMNQGA][LIVMNQGA SEQ ID NO:667)]-x(2)-[LIVMFT][LIVMFT SEQ ID NO:282)]-[GSTANC][GSTANC SEQ ID NO:668)]-[LIVMFYWSTAC][LIVMFYWSTAC SEQ ID NO:669)]-[DENH][DENH SEQ ID NO:670)]-R-[FYWCSH][FYWCSH SEQ ID NO:671)]-x(2)-[LIVM][LIVM SEQ ID NO:4)]

[1] Strosberg A.D.

Eur. J. Biochem. 196:1-10(1991).

[2] Kerlavage A.R.

Curr. Opin. Struct. Biol. 1:394-401(1991).

[3] Probst W.C., Snyder L.A., Schuster D.I., Brosius J., Sealfon S.C.

DNA Cell Biol. 11:1-20(1992).

[4] Savarese T.M., Fraser C.M.

Biochem. J. 283:1-9(1992).

[5] Branchek T.

Curr. Biol. 3:315-317(1993).

[6] Stiles G.L.

J. Biol. Chem. 267:6451-6454(1992).

[7] Friell T., Kobilka B.K., Lefkowitz R.J., Caron M.G.

Trends Neurosci. 11:321-324(1988).

[8] Stevens C.F.

Curr. Biol. 1:20-22(1991).

[9] Sakurai T., Yanagisawa M., Masaki T.

Trends Pharmacol. Sci. 13:103-107(1992).

[10] Salesse R., Remy J.J., Levin J.M., Jallal B., Garnier J.

5 Biochimie 73:109-120(1991).

[11] Lancet D., Ben-Arie N.

Curr. Biol. 3:668-674(1993).

[12] Uhl G.R., Childers S., Pasternak G.

Trends Neurosci. 17:89-93(1994).

10 [13] Barnard E.A., Burnstock G., Webb T.E.

Trends Pharmacol. Sci. 15:67-70(1994).

[14] Applebury M.L., Hargrave P.A.

Vision Res. 26:1881-1895(1986).

[15] Attwood T.K., Eliopoulos E.E., Findlay J.B.C.

15 Gene 98:153-159(1991).

(7tm 1) Visual pigments (opsins) retinal binding site

Visual pigments [1,2] are the light-absorbing molecules that mediate vision.

They consist of an apoprotein, opsin, covalently linked to the chromophore

20 cis-retinal. Vision is effected through the absorption of a photon by cis-

retinal which is isomerized to trans-retinal. This isomerization leads to a

change of conformation of the protein. Opsins are integral membrane proteins with seven transmembrane regions that belong to family 1 of G-protein coupled receptors.

25

In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).

30

In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6,

R7 and R8) expresses a specific opsin.

Proteins evolutionary related to opsins include squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal and mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern that had been developed includes this residue.

-Consensus pattern: [LIVMWAC][LIVMWAC SEQ ID NO:672)]-[PGC]-x(3)-[SAC]-K-[STALIMR][STALIMR SEQ ID NO:673)]-[GSACPNV][GSACPNV SEQ ID NO:674)]-[STACP][STACP SEQ ID NO:384)]-x(2)-[DENF][DENF SEQ ID NO:675)]-[AP]-x(2)-[IY]
[K is the retinal binding site]

[1] Applebury M.L., Hargrave P.A.
Vision Res. 26:1881-1895(1986).

[2] Fryxell K.J., Meyerowitz E.M.
J. Mol. Evol. 33:367-378(1991).

[3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
Biochemistry 33:13117-13125(1994).

The following descriptions of protein family functions are not provided by the Pfam or Prosite databases.

740. BAH

BAH domain. Number of members: 65

[1] Medline: 97074677. Molecular cloning of polybromo, a nuclear protein containing multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain. Nicolas RH, Goodwin GH; Gene 1996;175:233-240.

[2] Medline: 99198739. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. Callebaut I, Courvalin J-C, Mornon JP; FEBS letts 1999;446:189-193.

741. ELM2.

ELM2 domain. The ELM2 (Egl-27 and MTA1 homology 2) domain is a small domain of unknown function. Number of members: 10

742. Euk proin. EUKARYOTIC_PORIN The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV.

This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2); vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5].

A conserved region located at the C-terminal part of these proteins was selected as a signature pattern.

Consensus pattern[YH]-x(2)-D-[SPCAD][SPCAD SEQ ID NO:676]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF][LIVMF SEQ ID NO:2]-[DNSTA][DNSTA SEQ ID NO:677]-[DNS]-x(4)-[GSTAN][GSTAN SEQ ID NO:296]-[LIVMA][LIVMA SEQ ID NO:30]-x-[LIVMY][LIVMY SEQ ID NO:141]

[1] Benz R. Biochim. Biophys. Acta 1197:167-196(1994).

[2] Manella C.A. Trends Biochem. Sci. 17:315-320(1992).

[3] Dihanich M. Experientia 46:146-153(1990).

[4] Forte M., Guy H.R., Mannella C.A. J. Bioenerg. Biomembr. 19:341-350(1987).

[5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W.J. Genomics 36:192-196(1996).

5 743. Glyco hydor 19

Chitinases family 19 signatures

cross-reference(s) CHITINASE_19_1, CHITINASE_19_2

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence
 10 similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II) are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant
 15 entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues.

Two highly conserved regions were selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

20

Consensus pattern C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF][LIVMF SEQ ID NO:2]-x-A-x(3)-[YF]-x(2)-F-[GSA]

Consensus pattern [LIVM][LIVM SEQ ID NO:4]-[GSA]-F-x-[STAG][STAG SEQ ID NO:20]](2)-[LIVMFY][LIVMFY SEQ ID NO:18]]-W-[FY]-W-[LIVM][LIVM SEQ ID NO:4]

25

[1]Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).

[2] Henrissat B. Biochem. J. 280:309-316(1991).

30

744. MBD

Methyl-CpG binding domain

The Methyl-CpG binding domain (MBD) binds to DNA that contains one or more symmetrically methylated CpGs [1]. DNA methylation in animals is associated with alterations in chromatin structure and silencing of gene expression. MBD has negligible non-specific affinity for DNA. In vitro foot-printing with MeCP2 showed the MBD can protect a 12 nucleotide region surrounding a methyl CpG pair [1]. MBDs are found in several Methyl-CpG binding proteins and also DNA demethylase [2]. Number of members: 11

[1]Medline: 94232813. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nan X, Meehan RR, Bird A; Nucleic Acids Res 1993;21:4886-4892.

[2]Medline: 99158138. A mammalian protein with specific demethylase activity for mCpG DNA. Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M; Nature 1999;397:579-583.

745. Peptidase C1

Eukaryotic thiol (cysteine) proteases active sites

cross-reference(s) THIOL_PROTEASE_CYS; THIOL_PROTEASE_HIS;
THIOL_PROTEASE_ASN

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].

- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].

- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium-activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain.

- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].

- Human cathepsin O [4].

- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).

617

- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; 5 rice oryzain alpha, beta, and gamma; tomato low-temperature induced, *Arabidopsis thaliana* A494, RD19A and RD21A.
- House-dust mites allergens DerP1 and EurM1.
- Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen 10 SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3).
- Slime mold cysteine proteinases CP1 and CP2.
- Cruzipain from *Trypanosoma cruzi* and *brucei*.
- Trophozoite cysteine proteinase (TCP) from various *Plasmodium* species.
- 15 - Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*.
- Baculoviruses cathepsin-like enzyme (v-cath).
- *Drosophila* small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- 20 - *Caenorhabditis elegans* hypothetical protein C06G4.2, a calpain-like protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from *Lactococcus lactis* (gene pepC) [5].
- 25 - Thiol protease tpr from *Porphyromonas gingivalis*.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

- 30 - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.

- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <PDOC00382>).

- Plasmodium falciparum serine-repeat protein (SERA), the major blood stage antigen.

- 5 This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

The sequences around the three active site residues are well conserved and can be used as signature patterns.

- 10 Consensus pattern Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC][STAGC SEQ ID NO:45)]-[STAGCV][STAGCV SEQ ID NO:159)] [C is the active site residue]

Note the residue in position 4 of the pattern is almost always cysteine; the only exceptions are calpains (Leu), bleomycin hydrolase (Ser) and yeast YCP1 (Ser). Note the residue in position

- 15 5 of the pattern is always Gly except in papaya protease IV where it is Glu.

Consensus pattern [LIVMGSTAN][LIVMGSTAN SEQ ID NO:160)]-x-H-[GSACE][GSACE SEQ ID NO:161)]-[LIVM][LIVM SEQ ID NO:4)]-x-[LIVMAT][LIVMAT SEQ ID NO:162)](2)-G-x-[GSADNH][GSADNH SEQ ID NO:163)] [H is the active site residue]

- 20 Consensus pattern [FYCH][FYCH SEQ ID NO:164)]-[WI]-[LIVT][LIVT SEQ ID NO:165)]-x-[KRQAG][KRQAG SEQ ID NO:166)]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW][LFYW SEQ ID NO:167)]-[LIVMFYG][LIVMFYG SEQ ID NO:168)]-x-[LIVMF][LIVMF SEQ ID NO:2)] [N is the active site residue]

Note these proteins belong to family C1 (papain-type) and C2 (calpains) in the classification of peptidases [7,E1].

25

[1]Dufour E. Biochimie 70:1335-1342(1988).

[2]Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3]Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

- 30 [4]Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

[5]Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6]Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[7]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

5 746. Peptidase M22

Glycoprotease family signature cross-reference(s) GLYCOPROTEASE

Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-sialoglycoprotein endopeptidase, is a metalloprotease secreted by *Pasteurella haemolytica* which specifically cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

- *Escherichia coli* hypothetical protein ygjD (ORF-X).
- *Bacillus subtilis* hypothetical protein ydiE.
- *Mycobacterium leprae* hypothetical protein U229E.
- 15 - *Mycobacterium tuberculosis* hypothetical protein MtCY78.10.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0807.
- *Methanococcus jannaschii* hypothetical protein MJ1130.
- *Haloarcula marismortui* hypothetical protein in HSH 3'region.
- Yeast hypothetical protein YKR038c.
- 20 - Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

25 Consensus pattern[KR]-[GSAT][GSAT SEQ ID NO:100]]-x(4)-[FYWLH][FYWLH SEQ ID NO:273]]-[DQNGK][DQNGK SEQ ID NO:274]]-x-P-x-[LIVMFY][LIVMFY SEQ ID NO:18]]-x(3)-H-x(2)-[AG]-H-[LIVM][LIVM SEQ ID NO:4]]

Note these proteins belong to family M22 in the classification of peptidases [2,E1].

[1]Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).

[2]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

747. SAM. SAM domain (Sterile alpha motif)

It has been suggested that SAM is an evolutionarily conserved protein binding domain that is involved in the regulation of numerous developmental processes in diverse eukaryotes. The SAM domain can potentially function as a protein interaction module through its ability to homo- and heterooligomerise with other SAM domains. Number of members: 81

[1]Medline: 96100659 SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins Ponting CP; Prot Sci 1995;4:1928-1930.

[2]Medline: 97160498 SAM as a protein interaction domain involved in developmental regulation. Shultz J, Ponting CP, Hofmann K, Bork P; Prot Sci 1997;6:249-253.

[3]Medline: 99101382 The crystal structure of an Eph receptor SAM domain reveals a mechanism for modular dimerization. Reference Author: Stapleton D, Balan I, Pawson T, Sicheri F; Nat Struct Biol 1999;6:44-49.

748. Tyrosinase signatures cross-reference(s) TYROSINASE_1; TYROSINASE_2

Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols. This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the formation of pigments such as melanins and other polyphenolic compounds.

Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].

At least two proteins related to tyrosinase are known to exist in mammals:

- TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydroxyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid.

- TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPachrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPachrome to DHICA. TRP-2 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7].

5

Other proteins that belong to this family are:

- Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8].

10 - Caenorhabditis elegans hypothetical protein C02C2.1.

Two signature patterns for tyrosinase and related proteins have been derived. The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.

15

Consensus pattern H-x(4,5)-F-[LIVMFTP][LIVMFTP SEQ ID NO:678]-x-[FW]-H-R-x(2)-[LM]-x(3)-E

[The two H's are copper ligands]

20 Consensus pattern D-P-x-F-[LIVMFYW][LIVMFYW SEQ ID NO:26]-x(2)-H-x(3)-D [H is a copper ligand]

[1] Lerch K. Prog. Clin. Biol. Res. 256:85-98(1988).

25 [2] Jackman M.P., Hajnal A., Lerch K. Biochem. J. 274:707-713(1991).

[3] Linzen B. Naturwissenschaften 76:206-211(1989).

[4] Lang W.H., van Holde K.E. Proc. Natl. Acad. Sci. U.S.A. 88:244-248(1991).

[5] Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borrón J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).

30 [6] Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).

[7] Solano F., Martínez-Liarte J.H., Jimenez-Cervantes C., Garcia-Borrón J.C., Lozano J.A. Biochem. Biophys. Res. Commun. 204:1243-1250(1994).

[8]Cary J.W., Lax A.R., Flurkey W.H. Plant Mol. Biol. 20:245-253(1992).

749. (Mur Ligase) Folylpolyglutamate synthase signatures

- 5 Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in
10 glycine residues and could play a role in the catalytical activity and/or in substrate binding.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVMFY][LIVMFY SEQ ID NO:18]-x-[LIVM][LIVM SEQ ID NO:4]-
15 [STAG][STAG SEQ ID NO:20]-G-T-[NK]-G-K-x-[ST]-x(7)- [LIVM][LIVM SEQ ID NO:4]](2)-x(3)-[GSK]

Consensus pattern[LIVMFY][LIVMFY SEQ ID NO:18]](2)-E-x-G-[LIVM][LIVM SEQ ID NO:4]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM][LIVM SEQ ID NO:4]](2)

- 20 [1]Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

750. (Peptidase M3) Neutral zinc metallopeptidases, zinc-binding region signature

- 25 The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the
30 proteases which are currently known to belong to these families.

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).

623

- Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- 5 - Yeast alanine/arginine aminopeptidase (gene AAP1).
- Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

10

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

15

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).
- 20 - Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).
- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- 25 - Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).
- Yeast hypothetical protein YKL134c.

Family M4

- 30 - Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus.
- Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB).
- Extracellular elastase from Staphylococcus epidermidis.

- Extracellular protease prt1 from *Erwinia carotovora*.
- Extracellular minor protease smp from *Serratia marcescens*.
- Vibriolysin (EC 3.4.24.25) from various species of *Vibrio*.
- Protease prtA from *Listeria monocytogenes*.
- 5 - Extracellular proteinase proA from *Legionella pneumophila*.

Family M5

- Mycolysin (EC 3.4.24.31) from *Streptomyces cacaoi*.

10 Family M6

- Immune inhibitor A from *Bacillus thuringiensis* (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- 15 - *Streptomyces* extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of *Leishmania*.

20

Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

25 Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.
- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene aprA).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.
- Yeast hypothetical protein YIL108w.

30

Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC

625

3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylsin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).

- 5 - Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
 - Soybean metalloendoprotease 1.

Family M11

- 10 - *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
 - Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border
15 metalloendopeptidase.
 - Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog of BMP-1 is the dorsal-ventral patterning protein tolloid.
 - Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN
20 from *Strongylocentrotus purpuratus*.
 - *Caenorhabditis elegans* protein toh-2.
 - *Caenorhabditis elegans* hypothetical protein F42A10.8.
 - Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown
25 of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in
30 hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimere lysin I (EC 3.4.25.52) and II (EC 3.4.25.53).
 - Mouse cell surface antigen MS2.

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

Family M35

- Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

Description of pattern(s) and/or profile(s)

10 Consensus pattern[GSTALIVN][GSTALIVN SEQ ID NO:679]-x(2)-H-E-
[LIVMFYW][LIVMFYW SEQ ID NO:26)]-{DEHRKP}-{DEHRKP SEQ ID NO:680)}-H-x-
[LIVMFYWGSPQ][LIVMFYWGSPQ SEQ ID NO:681)] [The
two H's are zinc ligands] [E is the active site residue]

Sequences known to belong to this class detected by the patternALL,
15 except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT55; including Neurospora
crassa conidiation-specific protein 13 which could be a
zinc-protease.

- [1]Jongeneel C.V., Bouvier J., Bairoch A.
20 FEBS Lett. 242:211-214(1989).
- [2]Murphy G.J.P., Murphy G., Reynolds J.J.
FEBS Lett. 289:4-7(1991).
- [3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay
D.B., Stoecker W.
25 Zoology 99:237-246(1996).
- [4]Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).
- [5]Woessner J. Jr.
FASEB J. 5:2145-2154(1991).
- [6]Hite L.A., Fox J.W., Bjarnason J.B.
30 [7]Montecucco C., Schiavo G.
Trends Biochem. Sci. 18:324-327(1993).
- [8]Niemann H., Blasi J., Jahn R.

Trends Cell Biol. 4:179-185(1994).

751. PseudoU_synt_1

- 5 tRNA pseudouridine synthase is involved in the formation of pseudouridine at the anticodon stem and loop of transfer-RNAs Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl) uracil, and is the most abundant modified nucleoside found in all cellular RNAs. The TruA-like proteins also exhibit a conserved sequence with a strictly conserved aspartic acid, likely involved in catalysis. Number of members: 25

10

[1]Medline: 98254513. Transfer RNA-pseudouridine synthetase Pus1 of *Saccharomyces cerevisiae* contains one atom of zinc essential for its native conformation and tRNA recognition. Arluison V, Hountondji C, Robert B, Grosjean H; *Biochemistry* 1998;37:7268-7276.

15

752. EPSP synthase signatures

- 20 EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase) (EC 2.5.1.19) catalyzes the sixth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene *aroA*), plants and fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway) [1]. EPSP synthase has been extensively studied as it is the target of the potent herbicide glyphosate which inhibits the enzyme.

- 25 The sequence of EPSP from various biological sources shows that the structure of the enzyme has been well conserved throughout evolution. Two conserved regions were selected as signature patterns. The first pattern corresponds to a region that is part of the active site and which is also important for the resistance to glyphosate [2]. The second pattern is located in the C-terminal part of the protein and contains a conserved lysine which seems to be
- 30 important for the activity of the enzyme.

Description of pattern(s) and/or profile(s)

629

Consensus pattern[LIVM][LIVM SEQ ID NO:4)]-x(2)-[GN]-N-[SA]-G-T-[STA]-x-R-x-
[LIVMY][LIVMY SEQ ID NO:141)]-x-[GSTA][GSTA SEQ ID NO:19)]

Consensus pattern[KR]-x-[KH]-E-[CST]-[DNE]-R-[LIVM][LIVM SEQ ID NO:4)]-x-[STA]-
[LIVMC][LIVMC SEQ ID NO:142)]-x(2)-[EN]-[LIVMF][LIVMF SEQ ID NO:2)]-x-
5 [KRA]-[LIVMF][LIVMF SEQ ID NO:2)]-G

[1]Stallings W.C., Abdel-Megid S.S., Lim L.W., Shieh H.-S., Dayringer H.E., Leimgruber
N.K., Stegeman R.A., Anderson K.S., Sikorski J.A., Padgett S.R., Kishore G.M. Proc.
Natl. Acad. Sci. U.S.A. 88:5046-5050(1991).

10 [2]Padgett S.R., Re D.B., Gaser C.S., Eicholtz D.A., Frazier R.B., Hironaka C.M., Levine
E.B., Shah D.M., Fraley R.T., Kishore G.M. J. Biol. Chem. 266:22364-22369(1991).

753. Glyco_hydro_18

15 Glycosyl hydrolases family 18. Number of members: 173

[1]Medline: 95219379. Crystal structure of a bacterial chitinase at 2.3 Å resolution. Perrakis
A, Tews I, Dauter Z, Oppenheim AB, Chet I, Wilson KS, Vorgias CE; Structure
1994;2:1169-1180.

20

754. Esterase

Putative esterase

This family contains Esterase D Swiss:P10768. However it is not clear if all members of the
family have the same function. This family is possibly related to the COesterase family.

25 Number of members: 36

755. (HMA) Heavy-metal-associated domain

A conserved domain of about 30 amino acid residues has been found [1] in a number of
30 proteins that transport or detoxify heavy metals. This domain contains two conserved
cysteines that could be involved in the binding of these metals. The domain has been
termed Heavy-Metal-Associated (HMA). It has been found in:

- A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPases ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from Enterococcus faecalis and synA from Synechococcus contain one copy of the HMA domain. The cadmium ATPases cadA from Bacillus firmus and from plasmid pI258 from Staphylococcus aureus also contain a single HMA domain, while a chromosomal Staphylococcus aureus cadA contains two copies. Other, less characterized ATPases that contain the HMA domain are: fixI from Rhizobium meliloti, pacS from Synechococcus strain PCC 7942), Mycobacterium leprae ctpA and ctpB and Escherichia coli hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section.
- Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from Bacillus strain RC607 which has two) in the N-terminal part of merA.
- Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain.
- Helicobacter pylori copper-binding protein copP.
- Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper.

The consensus pattern for HMA spans the complete domain.

Description of pattern(s) and/or profile(s)

30 Consensus pattern [LIVN][LIVN SEQ ID NO:682]]-x(2)-[LIVMFA][LIVMFA SEQ ID NO:81]]-x-C-x-[STAGCDNH][STAGCDNH SEQ ID NO:683]]-C-x(3)-[LIVFG][LIVFG SEQ ID NO:684]]-x(3)-[LIV]-x(9,11)-[IVA]-x-[LVFYYS][LVFYYS SEQ ID NO:685]] [The two C's probably bind metals]

[1]Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994).

[2]Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995).

5 756. (Peptidase M10) Matrixins cysteine switch

PROSITE cross-reference(s): CYSTEINE_SWITCH

Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins

[1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an
 10 inactive form (zymogen) that differs from the mature enzyme by the presence of an N-
 terminal propeptide. A highly conserved octapeptide is found two residues downstream of
 the C-terminal end of the propeptide. This region has been shown to be involved in
 autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site
 zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or
 'autoinhibitor region'.

15 A cysteine switch has been found in the following zinc proteases:

- MMP-1 (EC 3.4.24.7) (interstitial collagenase).
- MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).
- MMP-3 (EC 3.4.24.17) (stromelysin-1).
- 20 - MMP-7 (EC 3.4.24.23) (matrilysin).
- MMP-8 (EC 3.4.24.34) (neutrophil collagenase).
- MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).
- MMP-10 (EC 3.4.24.22) (stromelysin-2).
- MMP-11 (EC 3.4.24.-) (stromelysin-3).
- 25 - MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- MMP-13 (EC 3.4.24.-) (collagenase 3).
- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 1).
- MMP-15 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 2).
- MMP-16 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 3).
- 30 - Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].
- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

Description of pattern(s) and/or profile(s)

Consensus pattern P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ][LIVSAPKQ SEQ ID NO:372] [C
chelates the zinc ion]

- [1]Woessner J. Jr. FASEB J. 5:2145-2154(1991).
- 5 [2]Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).
- [3]Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).
- [4]Lepage T., Gache C. EMBO J. 9:3003-3012(1990).
- 10 [5]Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

757. (Peptidase S8) Serine proteases, subtilase family, active sites

- 15 PROSITE cross-reference(s): PS00136; SUBTILASE_ASP, PS00137; SUBTILASE_HIS, PS00138; SUBTILASE_SER

Subtilases [1,2] are an extensive family of serine proteases whose catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the
20 residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.

The subtilase family currently includes the following proteases:

- Subtilisins (EC 3.4.21.62), these alkaline proteases from various *Bacillus* species have
25 been the target of numerous studies in the past thirty years.
- Alkaline elastase YaB from *Bacillus* sp. (gene ale).
- Alkaline serine exoprotease A from *Vibrio alginolyticus* (gene proA).
- Aqualysin I from *Thermus aquaticus* (gene pstI).
- AspA from *Aeromonas salmonicida*.
- 30 - Bacillopeptidase F (esterase) from *Bacillus subtilis* (gene bpf).
- C5A peptidase from *Streptococcus pyogenes* (gene scpA).
- Cell envelope-located proteases PI, PII, and PIII from *Lactococcus lactis*.
- Extracellular serine protease from *Serratia marcescens*.

- Extracellular protease from *Xanthomonas campestris*.
- Intracellular serine protease (ISP) from various *Bacillus*.
- Minor extracellular serine protease epr from *Bacillus subtilis* (gene epr).
- Minor extracellular serine protease vpr from *Bacillus subtilis* (gene vpr).
- 5 - Nisin leader peptide processing protease nisP from *Lactococcus lactis*.
- Serotype-specific antigene 1 from *Pasteurella haemolytica* (gene ssal).
- Thermitase (EC 3.4.21.66) from *Thermoactinomyces vulgaris*.
- Calcium-dependent protease from *Anabaena variabilis* (gene prcA).
- Halolysin from halophilic bacteria sp. 172p1 (gene hly).
- 10 - Alkaline extracellular protease (AEP) from *Yarrowia lipolytica* (gene xpr2).
- Alkaline proteinase from *Cephalosporium acremonium* (gene alp).
- Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1).
- Cuticle-degrading protease (pr1) from *Metarhizium anisopliae*.
- KEX-1 protease from *Kluyveromyces lactis*.
- 15 - Kexin (EC 3.4.21.61) from yeast (gene KEX-2).
- Oryzin (EC 3.4.21.63) (alkaline proteinase) from *Aspergillus* (gene alp).
- Proteinase K (EC 3.4.21.64) from *Tritirachium album* (gene proK).
- Proteinase R from *Tritirachium album* (gene proR).
- Proteinase T from *Tritirachium album* (gene proT).
- 20 - Subtilisin-like protease III from yeast (gene YSP3).
- Thermomycin (EC 3.4.21.65) from *Malbranchea sulfurea*.
- Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4 protease from mammals, other vertebrates, and invertebrates. These proteases are involved in the processing of hormone precursors at sites comprised of pairs of basic amino acid residues [3].
- 25 - Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human.
- Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain (see <PDOC00185>).

30

Description of pattern(s) and/or profile(s)

634

Consensus pattern[STAIV][STAIV SEQ ID NO:130]-x-[LIVMF][LIVMF SEQ ID NO:2]-
[LIVM][LIVM SEQ ID NO:4]-D-[DSTA][DSTA SEQ ID NO:686]-G-

[LIVMFC][LIVMFC SEQ ID NO:90]-x(2,3)-[DNH] [D is the active site residue]

Consensus patternH-G-[STM]-x-[VIC]-[STAGC][STAGC SEQ ID NO:45]-[GS]-x-

[LIVMA][LIVMA SEQ ID NO:30]-[STAGCLV][STAGCLV SEQ ID NO:687]-

[SAGM][SAGM SEQ ID NO:688] [H is the active site residue]

Consensus patternG-T-S-x-[SA]-x-P-x(2)-[STAVC][STAVC SEQ ID NO:505]-[AG] [S is
the active site residue]

Note if a protein includes at least two of the three active site signatures, the probability of it
being a serine protease from the subtilase family is 100%

Note these proteins belong to family S8 in the classification of
peptidases [5,E1].

[1]Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4:719-
737(1991).

[2]Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).

[3]Barr P.J. Cell 66:1-3(1991).

[4]Shaulsky G., Kuspa A., Loomis W.F.; Genes Dev. 9:1111-1122(1995).

[5]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

758. (SSB) Single-strand binding protein family signatures

PROSITE cross-reference(s): PS00735; SSB_1,PS00736; SSB_2

The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-
destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to
single-stranded DNA (ss-DNA) and plays an important role in DNA replication,
recombination and repair.

Closely related variants of SSB are encoded in the genome of a variety of large self-
transmissible plasmids. SSB has also been characterized in bacteria such as Proteus mirabilis
or Serratia marcescens.

Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in mitochondrial DNA replication are structurally and evolutionary related to prokaryotic SSB. Proteins currently known to belong to this subfamily are listed below [2].

- Mammalian protein Mt-SSB (P16).
- 5 - Xenopus Mt-SSBs and Mt-SSBr.
- Drosophila MtSSB.
- Yeast protein RIM1.

Two signature patterns have been developed for these proteins. The first is a conserved
10 region in the N-terminal section of the SSB's. The second is a centrally located region which, in Escherichia coli SSB, is known to be involved in the binding of DNA.

Description of pattern(s) and/or profile(s)

Consensus pattern [LIVMF][LIVMF SEQ ID NO:2)]-[NST]-[KRT]-[LIVM][LIVM SEQ ID
15 NO:4)]-x-[LIVMF][LIVMF SEQ ID NO:2)](2)-G-[NHRK][NHRK SEQ ID NO:689)]-
[LIVM][LIVM SEQ ID NO:4)]- [GST]-x-[DET]

Consensus pattern T-x-W-[HY]-[RNS]-[LIVM][LIVM SEQ ID NO:4)]-x-[LIVMF][LIVMF
SEQ ID NO:2)]-[FY]-[NGKR][NGKR SEQ ID NO:690)]

- 20 [1]Meyer R.R., Laine P.S. Microbiol. Rev. 54:342-380(1990).
[2]Stroumbakis N.D., Li Z., Tolia P.P. Gene 143:171-177(1994).

759. KDPG and KHG aldolases active site signatures

PROSITE cross-reference(s): PS00159; ALDOLASE_KDPG_KHG_1, PS00160;
25 ALDOLASE_KDPG_KHG_2

4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the
interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-
dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the
30 interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and
glyceraldehyde 3-phosphate.

These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.

Two signature patterns were developed for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.

Description of pattern(s) and/or profile(s)

Consensus pattern G-[LIVM][LIVM SEQ ID NO:4]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]

Consensus pattern G-x(3)-[LIVMF][LIVMF SEQ ID NO:2]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]

[1] Vlahos C J., Dekker E.E. J. Biol. Chem. 263:11683-11691(1988).

760. AP endonucleases family 1 signatures. PROSITE cross-reference(s): PS00726; AP_NUCLEASE_F1_1, PS00727; AP_NUCLEASE_F1_2, PS00728; AP_NUCLEASE_F1_3

DNA damaging agents such as the antitumor drugs bleomycin and neocarzinostatin or those that generate oxygen radicals produce a variety of lesions in DNA. Amongst these is base-loss which forms apurinic/apyrimidinic (AP) sites or strand breaks with atypical 3'termini.

DNA repair at the AP sites is initiated by specific endonuclease cleavage of the phosphodiester backbone. Such endonucleases are also generally capable of removing blocking groups from the 3'terminus of DNA strand breaks.

AP endonucleases can be classified into two families on the basis of sequence similarity.

Family 1 groups the enzymes listed below [1].

- Escherichia coli exonuclease III (EC 3.1.11.2) (gene xthA).
- Streptococcus pneumoniae and Bacillus subtilis exonuclease A (gene exoA).

637

- Mammalian AP endonuclease 1 (AP1) (EC 4.2.99.18).
- Drosophila recombination repair protein 1 (gene Rrp1).
- Arabidopsis thaliana apurinic endonuclease-redox protein (gene arp).

5 Except for Rrp1 and arp, these enzymes are proteins of about 300 amino-acid residues. Rrp1 and arp both contain additional and unrelated sequences in their N-terminal section (about 400 residues for Rrp1 and 270 for arp).

10 Three signature patterns were developed for this family of enzymes. The patterns are based on the most conserved regions. The first pattern contains a glutamate which has been shown [2], in the Escherichia coli enzyme to bind a divalent metal ion such as magnesium or manganese

15 Consensus pattern[APF]-D-[LIVMF][LIVMF SEQ ID NO:2](2)-x-[LIVM][LIVM SEQ ID NO:4]-Q-E-x-K [E binds a divalent metal ion]

Consensus patternD-[ST]-[FY]-R-[KH]-x(7,8)-[FYW]-[ST]-[FYW](2)

Consensus patternN-x-G-x-R-[LIVM][LIVM SEQ ID NO:4]-D-[LIVMFYH][LIVMFYH SEQ ID NO:541]-x-[LV]-x-S

20 [1] Barzilay G., Hickson I.S. BioEssays 17:713-719(1995).

[2] Mol C.D., Kuo C.-F., Thayer M.M., Cunningham R.P., Tainer J.A. Nature 374:381-386(1995).

25 761. (ER)Enhancer of rudimentary signature, PROSITE cross-reference(s): PS01290; ER

The Drosophila protein 'enhancer of rudimentary' (gene (e(r))) is a small protein of 104 residues whose function is not yet clear. From an evolutionary point of view, it is highly conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle.

30

A conserved region in the central part of the protein was selected as as signaure pattern.

Consensus patternY-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S

[1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).

- 5 762. (ETF alpha) Electron transfer flavoprotein alpha-subunit signature, PROSITE cross-reference(s): PS00696; ETF_ALPHA

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory
10 chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.

The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the
15 bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.

Other related proteins are:

- 20 - Escherichia coli hypothetical protein ydiR.
- Escherichia coli hypothetical protein ygcQ.

A highly conserved region which is located in the C-terminal section was selected as a signature pattern for these proteins.

25

Consensus pattern [LI]-Y-[LIVM][LIVM SEQ ID NO:4]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N

[1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).

30 [2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

763. (lectin c) C-type lectin domain signature and profile

- A number of proteins expressed on the surface of natural killer T-cells: NKG2, NKR-P1, YE1/88 (Ly-49), CD69 and on B-cells: CD72, LyB-2. The CTL-domain in these proteins is distantly related to other CTL-domains; it is unclear whether they are likely to bind carbohydrates.

5

Proteins that consist of an N-terminal collagenous domain followed by a CTL-domain [5], these proteins are sometimes called 'collectins':

- Pulmonary surfactant-associated protein A (SP-A). SP-A is a calcium-dependent protein that binds to surfactant phospholipids and contributes to lower the surface tension at the air-liquid interface in the alveoli of the mammalian lung.

10

- Pulmonary surfactant-associated protein D (SP-D).

- Conglutinin, a calcium-dependent lectin-like protein which binds to a yeast cell wall extract and to immune complexes through the complement component (iC3b).

15

- Mannan-binding proteins (MBP) (also known as mannose-binding proteins). MBP's bind mannose and N-acetyl-D-glucosamine in a calcium-dependent manner.

20

- Bovine collectin-43 (CL-43).

Selectins (or LEC-CAM) [6,7]. Selectins are cell adhesion molecules implicated in the interaction of leukocytes with platelets or vascular endothelium. Structurally, selectins consist of a long extracellular domain, followed by a transmembrane region and a short cytoplasmic domain. The extracellular domain is itself composed of a CTL-domain, followed by an EGF-like domain and a variable number of SCR/Sushi repeats. Known selectins are:

25

- Lymph node homing receptor (also known as L-selectin, leukocyte adhesion molecule-1, (LAM-1), leu-8, gp90-mel, or LECAM-1)

30

- Endothelial leukocyte adhesion molecule 1 (ELAM-1, E-selectin or LECAM-2).

The ligand recognized by ELAM-1 is sialyl-Lewis x.

- Granule membrane protein 140 (GMP-140, P-selectin, PADGEM, CD62, or LECAM-

3). The ligand recognized by GMP-140 is Lewis x.

Large proteoglycans that contain a CTL-domain followed by one copy of a SCR/ Sushi repeat, in their C-terminal section:

5

- Aggrecan (cartilage-specific proteoglycan core protein). This proteoglycan is a major component of the extracellular matrix of cartilaginous tissues where it has a role in the resistance to compression.

- Brevican.

10

- Neurocan.

- Versican (large fibroblast proteoglycan), a large chondroitin sulfate proteoglycan that may play a role in intercellular signalling.

In addition to the CTL and Sushi domains, these proteins also contain, in their N-terminal domain, an Ig-like V-type region, two or four link domains (see <PDOC00955>) and up to two EGF-like repeats.

15

Two type-I membrane proteins:

20

- Mannose receptor from macrophages. This protein mediates the endocytosis of glycoproteins by macrophages in several recognition and uptake processes. Its extracellular section consists of a fibronectin type II domain followed by eight tandem repeats of the CTL domain.

25

- 180 Kd secretory phospholipase A2 receptor (PLA2-R). A protein whose structure is highly similar to that of the mannose receptor.

30

- DEC-205 receptor. This protein is used by dendritic cells and thymic epithelial cells to capture and endocytose diverse carbohydrate-binding antigens and direct them to antigen-processing cellular compartments. DEC-205 extracellular section consists of a fibronectin type II domain followed by ten tandem repeats of the CTL domain.

- Silk moth hemocytin, an humoral lectin which is involved in a self-defence mechanism. It is composed of 2 FA58C domains (see <PDOC00988>), a CTL domain, 2 VWFC domains (see <PDOC00928>), and a CTCK (see <PDOC00912>).

Various other proteins that uniquely consist of a CTL domain:

- 5 - Invertebrate soluble galactose-binding lectins. A category to which belong a humoral lectin from a flesh fly; echinoidin, a lectin from the coelomic fluid of a sea urchin; BRA-2 and BRA-3, two lectins from the coelomic fluid of a barnacle, a lectin from the tunicate *Polyandrocarpa misakiensis* and a newt oviduct lectin. The physiological importance of these lectins is not yet known but they may play an important role in defense mechanisms.
- 10 - Pancreatic stone protein (PSP) (also known as pancreatic thread protein (PTP), or reg), a protein that might act as an inhibitor of spontaneous calcium carbonate precipitation.
- Pancreatitis associated protein (PAP), a protein that might be involved in the control of bacterial proliferation.
- 15 - Tetranectin, a plasma protein that binds to plasminogen and to isolated kringle 4.
- Eosinophil granule major basic protein (MBP), a cytotoxic protein.
- A galactose specific lectin from a rattlesnake.
- Two subunits of a coagulation factor IX/factor X-binding protein (IX/X-bp),
- 20 a snake venom anticoagulant protein which binds with factors IX and X in the presence of calcium.
- Two subunits of a phospholipase A2 inhibitor from the plasma of a snake (PLI-A and PLI-B).
- A lipopolysaccharide-binding protein (LPS-BP) from the hemolymph of a
- 25 cockroach [8].
- Sea raven antifreeze protein (AFP) [9].

As a signature pattern for this domain, the C-terminal region with its three conserved cysteines was selected.

30

Consensus pattern C-[LIVMFYATG][LIVMFYATG SEQ ID NO:691]-x(5,12)-[WL]-x-[DNSR][DNSR SEQ ID NO:692]-x(2)-C-x(5,6)-

643

[FYWLIVSTA][FYWLIVSTA SEQ ID NO:693)][LIVMSTA][LIVMSTA SEQ ID NO:433)]-C [The three C's are involved in disulfide

bonds]

Note all CTL domains have five Trp residues before the second Cys,
5 with the exception of tunicate lectin and cockroach LPS-BP which have Leu.

Note this documentation entry is linked to both a signature pattern
and a profile. As the profile is much more sensitive than the
10 pattern, you should use it if you have access to the necessary software tools to do so.

[1] Drickamer K. J. Biol. Chem. 263:9557-9560(1988).

[2] Drickamer K. Prog. Nucleic Acid Res. Mol. Biol. 45:207-232(1993).

15 [3] Drickamer K. Curr. Opin. Struct. Biol. 3:393-400(1993).

[4] Spiess M. Biochemistry 29:10009-10018(1990).

[5] Weis W.I., Kahn R., Fourme R., Drickamer K., Hendrickson W.A. Science 254:1608-1615(1991).

[6] Siegelman M. Curr. Biol. 1:125-128(1991).

20 [7] Lasky L.A. Science 238:964-969(1992).

[8] Jomori T., Natori S. J. Biol. Chem. 266:13318-13323(1991).

[9] Ng N.F.L., Hew C.-L. J. Biol. Chem. 267:16069-16075(1992).

764. (SRCR) Speract receptor repeated domain signature

25 PROSITE cross-reference(s): PS00420; SPERACT_RECEPTOR,

The receptor for the sea urchin egg peptide speract is a transmembrane glycoprotein of 500 amino acid residues [1]. Structurally it consists of a large extracellular domain of 450 residues, followed by a transmembrane region and a small cytoplasmic domain of 12 amino
30 acids. The extracellular domain contains four repeats of a 115 amino acids domain. There are 17 positions that are perfectly conserved in the four repeats, among them are six cysteines, six glycines, and three glutamates.

Such a domain is also found, once, in the C-terminal section of mammalian macrophage scavenger receptor type I [2], a membrane glycoproteins implicated in the pathologic deposition of cholesterol in arterial walls during atherogenesis.

- 5 The signature pattern that was derived spans part of the N-terminal section of the domain and contains 8 of the 17 conserved residues.

Consensus pattern G-x(5)-G-x(2)-E-x(6)-W-G-x(2)-C-x(3)-[FYW]-x(8)-C-x(3)-G

- 10 [1] Dangott J.J., Jordan J.E., Bellet R.A., Garbers D.L. Proc. Natl. Acad. Sci. U.S.A. 86:2128-2132(1989).

[2] Freeman M., Ashkenas J., Rees D.J., Kingsley D.M., Copeland N.G., Jenkins N.A., Krieger M. Proc. Natl. Acad. Sci. U.S.A. 87:8810-8814(1990).

- 15 765. Bac_surface_Ag

Bacterial surface antigen

This entry includes the following surface antigens; D15 antigen from *H.influenzae*, OMA87 from *P.multocida*, OMP85 from *N.meningitidis* and *N.gonorrhoeae*. Number of members:

14

- 20

[1]Medline: 95255676. The sequencing of the 80-kDa D15 protective surface antigen of *Haemophilus influenzae*. Flack FS, Loosmore S, Chong P, Thomas WR; Gene 1995;156:97-99.

[2] Medline: 96333354. Cloning, sequencing, expression, and protective capacity of the oma87 gene encoding the *Pasteurella multocida* 87-kilodalton outer membrane antigen. Ruffolo CG, Adler B; Infect Immun 1996;64:3161-3167.

- 25

766. BRCA1 C Terminus (BRCT) domain

The BRCT domain is found predominantly in proteins involved in cell cycle checkpoint

- 30

functions responsive to DNA damage. It has been suggested that the Retinoblastoma protein contains a divergent BRCT domain, this has not been included in this family. The BRCT domain of XRCC1 forms a homodimer in the crystal structure Medline:99016060. This suggests that pairs of BRCT domains

associate as homo- or heterodimers. Number of members: 131

[1] Medline: 96259550. BRCA1 protein products ...Functional motifs... Koonin EV, Altschul SF, Bork P; Nature Genet 1996;13:266-268.

5 [2] Medline: 97153217. From BRCA1 to RAP1: A widespread BRCT module closely associated with DNA repair Callebaut I, Mornon JP; Febs lett 1997;400:25-30.

[3] Medline: 97186552. A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV; Faseb J 1997;11:68-76.

10 [4] Medline: 97402527. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ; Nucleic Acids Res 1997;25:3389-3402.

[5] Medline: 99016060. Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. Zhang X, Morera S, Bates PA, Whitehead PC, Coffey AI, Hainbucher K, 15 Nash RA, Sternberg MJ, Lindahl T, Freemont PS;

767. Kappa casein

Kappa-casein is a mammalian milk protein involved in a number of important physiological processes. In the gut, the ingested protein is split into an insoluble peptide (para kappa- 20 casein) and a soluble hydrophilic glycopeptide (caseinomacropeptide). Caseinomacropeptide is responsible for increased efficiency of digestion, prevention of neonate hypersensitivity to ingested proteins, and inhibition of gastric pathogens. Number of members: 56

[1] Medline: 98072500. Nucleotide sequence evolution at the kappa-casein locus: evidence 25 for positive selection within the family Bovidae. Ward TJ, Honeycutt RL, Derr JN; Genetics 1997;147:1863-1872.

768. Chitinases family 18 active site

PROSITE cross-reference(s) CHITINASE_18

30 Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl

hydrolases [2,E1]. Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:

a) Chitinases from:

- 5 - Prokaryotes such as *Alteromonas*, *Bacillus*, *Serratia*, *Streptomyces*, etc.
- Plants such as *Arabidopsis*, cucumber, bean, tobacco, etc.
- Fungi such as *Aphanocladium*, *Rhizopus*, *Saccharomyces*, etc.
- Nematode (*Brugia malayi*).
- Insects (*Manduca sexta*).
- 10 - Baculoviruses (*Autographa Californica* Nuclear Polyhedrosis virus).

b) Other proteins:

- Hevamine, a rubber tree protein with chitinase and lysozyme activities.
- 15 - *Kluyveromyces lactis* killer toxin alpha subunit, which acts as a chitinase.
- *Flavobacterium* and *Streptomyces* endo-beta-N-acetylglucosaminidases (EC 3.2.1.96).
- Mammalian di-N-acetylchitobiase which is involved in the degradation of asparagine-linked glycoproteins.
- Human cartilage glycoprotein Gp-39.
- 20 - Jack bean concanavalin B (conB), a protein that has lost its catalytic activity.

Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.

25 Consensus pattern [LIVMFY][LIVMFY SEQ ID NO:18]-[DN]-G-[LIVMF][LIVMF SEQ ID NO:2]-[DN]-[LIVMF][LIVMF SEQ ID NO:2]-[DN]-x-E [E is the active site residue]

[1] Flach J., Pilet P.-E., Jolles P. *Experientia* 48:701-716(1992).

30 [2] Henrissat B. *Biochem. J.* 280:309-316(1991).

[3] Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. *J. Biol. Chem.* 268:18567-18572(1993).

[4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).

[5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-1189(1994).

5

769. gag_p17. gag gene protein p17 (matrix protein).

The matrix protein forms an icosahedral shell associated with the inner membrane of the mature immunodeficiency virus. Number of members: 1598

10 [1] Medline: 95055757. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. Massiah MA, Starich MR, Paschall C, Summers MF, Christensen AM, Sundquist WI; J Mol Biol 1994;244:198-223.

770. GDA1/CD39 family of nucleoside phosphatases signature

15 PROSITE cross-reference(s); GDA1_CD39_NTPASE

A number of nucleoside diphosphate and triphosphate hydrolases as well as some yet uncharacterized proteins have been found to belong to the same family [1, 2]. This family currently consist of:

20 - Yeast guanosine-diphosphatase (EC 3.6.1.42) (GDPase) (gene GDA1). GDA1 is a golgi integral membrane enzyme that catalyzes the hydrolysis of GDP to GMP.

- Potato apyrase (EC 3.6.1.5) (adenosine diphosphatase) (ADPase). Apyrase acts on both ATP and ADP to produce AMP.

- Mammalian vascular ATP-diphosphohydrolase (EC 3.6.1.5) (also known as lymphoid cell activation antigen CD39).

25 - Toxoplasma gondii nucleoside-triphosphatases (EC 3.6.1.15) (NTPase). NTPase hydrolyses various nucleoside triphosphates to produce the corresponding nucleoside mono- and diphosphates. This enzyme is secreted into the invaded host cell into the parasitophorous vacuole, a specialized compartment where the parasite intracellularly resides.

30 - Pea nucleoside-triphosphatases (EC 3.6.1.15) (NTPase).

- Caenorhabditis elegans hypothetical protein C33H5.14.

- Caenorhabditis elegans hypothetical protein R07E4.4.

- Yeast chromosome V hypothetical protein YER005w.

The above uncharacterized proteins all seem to be membrane-bound.

- 5 All these proteins share a number of conserved domains. The best conserved of these domains have been selected. It is located in the central section of the proteins.

10 Consensus pattern[~~LIVM~~][LIVM SEQ ID NO:4]-x-G-x(2)-E-G-x-[FY]-x-[FW]-
[~~LIVA~~][LIVA SEQ ID NO:219]-[TAG]-x-N-[HY]

[1] Handa M., Guidotti G. Biochem. Biophys. Res. Commun. 218:916-923(1996).

[2] Vasconcelos E.G., Ferreira S.T., de Carvalho T.M.U., de Souza W., Kettlun A.M., Mancilla M., Valenzuela M.A., Verjovski-Almeida S. J. Biol. Chem. 271:22139-
15 22145(1996).

771. GTP cyclohydrolase I signatures

PROSITE cross-reference(s); GTP_CYCLOHYDROL_1_1, GTP_CYCLOHYDROL_1_2
GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and
20 dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-containing pigments in insects.

25 GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison of the sequence of the enzyme from bacterial and eukaryotic sources shows that the structure of this enzyme has been extremely well conserved throughout evolution [1].

30 Two conserved regions were selected as signature patterns. The first contains a perfectly conserved tetrapeptide which is part of the GTP-binding pocket [2], the second region also contains conserved residues involved in GTP-binding.

Consensus pattern[~~DEN~~]-[~~LIVM~~][LIVM SEQ ID NO:4](2)-x(2)-[~~KRNQ~~][KRNO SEQ ID NO:694]-[~~DEN~~]-[~~LIVM~~][LIVM SEQ ID NO:4]-x(3)-[ST]-x-C-E- H-H

Consensus pattern[SA]-x-[RK]-x-Q-[LIVM][LIVM SEQ ID NO:4]-Q-E-[RN]-[LI]-[TSN]

[1] Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H. Biochem. Biophys. Res. Commun. 212:705-711(1995).

- 5 [2] Nar H., Huber R., Meining W., Schmid C., Weinkauff S., Bacher A. Structure 3:459-466(1995).

772. IlvC. Acetohydroxy acid isomeroreductase

10 Acetohydroxy acid isomeroreductase catalyses the conversion of acetohydroxy acids into dihydroxy valerates. This reaction is the second in the synthetic pathway of the essential branched side chain amino acids valine and isoleucine. Number of members: 29

- [1] Medline: 97361822. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog
15 determined at 1.65 Å resolution. Biou V, Dumas R, Cohen-Addad C, Douce R, Job D, Pebay-Peyroula E; EMBO J 1997;16:3405-3415.

773. Prokaryotic membrane lipoprotein lipid attachment site

PROSITE cross-reference(s); PROKAR_LIPOPROTEIN

20 In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- 25 - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).
- Escherichia coli lipoprotein-34 (gene nlpB).
- Escherichia coli lipoprotein nlpC.
- Escherichia coli lipoprotein nlpD.
30 - Escherichia coli osmotically inducible lipoprotein B (gene osmB).
- Escherichia coli osmotically inducible lipoprotein E (gene osmE).
- Escherichia coli peptidoglycan-associated lipoprotein (gene pal).
- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).

- *Escherichia coli* copper homeostasis protein cutF (or nlpE).
 - *Escherichia coli* plasmids traT proteins.
 - *Escherichia coli* Col plasmids lysis proteins.
 - A number of *Bacillus* beta-lactamases.
 - 5 - *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA).
 - *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB).
 - *Borrelia hermsii* variable major protein 21 (gene vmp21) and 7 (gene vmp7).
 - *Chlamydia trachomatis* outer membrane protein 3 (gene omp3).
 - *Fibrobacter succinogenes* endoglucanase cel-3.
 - 10 - *Haemophilus influenzae* proteins Pal and Pcp.
 - *Klebsiella pullulunase* (gene pulA).
 - *Klebsiella pullulunase* secretion protein pulS.
 - *Mycoplasma hyorhinis* protein p37.
 - *Mycoplasma hyorhinis* variant surface antigens A, B, and C (genes vlpABC).
 - 15 - *Neisseria* outer membrane protein H.8.
 - *Pseudomonas aeruginosa* lipopeptide (gene lppL).
 - *Pseudomonas solanacearum* endoglucanase egl.
 - *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
 - *Rickettsia* 17 Kd antigen.
 - 20 - *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
 - *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
 - *Treponema pallidum* 34 Kd antigen.
 - *Treponema pallidum* membrane protein A (gene tmpA).
 - *Vibrio harveyi* chitobiase (gene chb).
 - 25 - *Yersinia* virulence plasmid protein yscJ.
- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper- binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).
- 30 From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.

651

Consensus pattern{DERK}{DERK SEQ ID NO:354}{(6)-

[LIVMFWSTAG][LIVMFWSTAG SEQ ID NO:352]}(2)-

[LIVMFYSTAGCQ][LIVMFYSTAGCQ SEQ ID NO:353]}-[AGS]-C [C is the lipid

attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the
 5 sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven
 positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2]Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

10 [3]von Heijne G. Protein Eng. 2:531-534(1989).

[4]Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol.
 Chem. 269:14939-14945(1994).

774. Aminoacyl-transfer RNA synthetases class-II signatures

15 PROSITE cross-reference(s); AA_TRNA_LIGASE_II_1; AA_TRNA_LIGASE_II_2

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate
 amino acids and transfer them to specific tRNA molecules as the first step in protein

biosynthesis. In prokaryotic organisms there are at least twenty different types of
 aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are

20 generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic
 form and a mitochondrial form. While all these enzymes have a common function, they are
 widely diverse in terms of subunit size and of quaternary structure.

The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine,

25 phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6]
 and probably have a common folding pattern in their catalytic domain for the binding of
 ATP and amino acid which is different to the Rossmann fold observed for the class I
 synthetases [7].

30 Class-II tRNA synthetases do not share a high degree of similarity, however at least three
 conserved regions are present [2,5,8]. Signature patterns from two of these regions have been
 derived.

652

Consensus pattern[FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern[GSTALVF][GSTALVF SEQ ID NO:42)]-~~{DENQHRKP}~~{DENQHRKP
SEQ ID NO:43)}-[GSTA][GSTA SEQ ID NO:19)]-[LIVMF][LIVMF SEQ ID NO:2)]-[DE]-
R-[LIVMF][LIVMF SEQ ID NO:2)]-x-[LIVMSTAG][LIVMSTAG SEQ ID NO:44)]-
5 [LIVMFY][LIVMFY SEQ ID NO:18)]

[1]Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[2]Delarue M., Moras D. BioEssays 15:675-687(1993).

[3]Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

10 [4]Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

[5]Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[6]Cusack S. Biochimie 75:1077-1081(1993).

[7]Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-
255(1990).

15 [8]Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

775. X. Trans-activation protein X

This protein is found in hepadnaviruses where it is indispensable for replication. Number of
members: 91

20

776. Thymidylate synthase active site

Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of
dUMP to dTMP with concomitant conversion of 5,10-methylenetetrahydrofolate to
dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an
25 important target for certain chemotherapeutic drugs.

Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in
protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate
reductase domain.

30

A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-
dihydro-dUMP intermediate). The sequence around the active site of this enzyme is
conserved from phages to vertebrates.

Consensus pattern R-x(2)-[LIVM][LIVM SEQ ID NO:4]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM][HAVM SEQ ID NO:695]-x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue]

- 5 [1] Benkovic S.J. Annu. Rev. Biochem. 49:227-251(1980).
[2] Ross P., O'Gara F., Condon S. Appl. Environ. Microbiol. 56:2156-2163(1990).

777. Glycosyl hydrolases family 31 signatures

10 It has been shown [1,2,3,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Lysosomal alpha-glucosidase (EC 3.2.1.20) (acid maltase) is a vertebrate glycosidase active at low pH, which hydrolyzes alpha(1->4) and alpha(1->6) linkages in glycogen, maltose, and isomaltose.
- Alpha-glucosidase (EC 3.2.1.20) from the yeast *Candida tsukunbaensis*.
- 15 - Alpha-glucosidase (EC 3.2.1.20) (gene malA) from the archebacteria *Sulfolobus solfataricus*.
- Intestinal sucrase-isomaltase (EC 3.2.1.48 / EC 3.2.1.10) is a vertebrate membrane-bound, multifunctional enzyme complex which hydrolyzes sucrose, maltose and isomaltose. The sucrase and isomaltase domains of the enzyme are homologous (41% of amino acid identity)
- 20 and have most probably evolved by duplication.
- Glucoamylase 1 (EC 3.2.1.3) (glucan 1,4-alpha-glucosidase) from various fungal species.
- Yeast hypothetical protein YBR229c.
- Fission yeast hypothetical protein SpAC30D11.01c.

25 An aspartic acid has been implicated [4] in the catalytic activity of sucrase, isomaltase, and lysosomal alpha-glucosidase. The region around this active residue is highly conserved and can be used as a signature pattern. A second region, which contains two conserved cysteines, has been used as an additional signature pattern.

30 Consensus pattern [GF]-[LIVMF][LIVMF SEQ ID NO:2]-W-x-D-M-[NSA]-E [D is the active site residue]

Consensus pattern G-[AV]-D-[LIVMTA][LIVMTA SEQ ID NO:311]-C-G-[FY]-x(3)-[ST]-x(3)-L-C-x-R-W-x(2)-[LV]-[GSA]-[SA]-F-x-P-F-x-R-[DN]

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Kinsella B.T., Hogan S., Larkin A., Cantwell B.A. Eur. J. Biochem. 202:657-664(1991).

[3] Naim H.Y., Niermann T., Kleinhans U., Hollenberg C.P., Strasser A.W.M. FEBS Lett. 294:109-112(1991).

5 [4] Hermans M.M.P., Kroos M.A., van Beeumen J., Oostra B.A., Reuser A.J.J. J. Biol. Chem. 266:13507-13512(1991).

778. Urease signatures

10 Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

15 Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

As signatures for this enzyme, a region was selected that contains two histidine that bind one of the nickel ions and the region of the active site histidine.

20 Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM][LIVM SEQ ID NO:4]-D-x-H-[LIVM][LIVM SEQ ID NO:4]-H-x(3)-P [The two H's bind nickel]

Consensus pattern [LIVM][LIVM SEQ ID NO:4](2)-[CT]-H-[HN]-L-x(3)-[LIVM][LIVM SEQ ID NO:4]-x(2)-D-[LIVM][LIVM SEQ ID NO:4]-x-F-A [H is the active site residue]

25 [1] Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

[2] Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

779. Tyrosine specific protein phosphatases signature and profiles

30 Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into

two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:

Soluble PTPases.

- 5 - PTPN1 (PTP-1B).
 - PTPN2 (T-cell PTPase; TC-PTP).
 - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.
 - 10 - PTPN5 (STEP).
 - PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The *Drosophila* protein corkscrew (gene *csw*) also belongs to this subgroup.
 - PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).
 - 15 - PTPN8 (70Z-PEP).
 - PTPN9 (MEG2).
 - PTPN12 (PTP-G1; PTP-P19).
 - Yeast PTP1.
 - Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.
 - 20 - Fission yeast *pyp1* and *pyp2* which play a role in inhibiting the onset of mitosis.
 - Fission yeast *pyp3* which contributes to the dephosphorylation of *cdc2*.
 - Yeast CDC14 which may be involved in chromosome segregation.
 - *Yersinia* virulence plasmid PTPases (gene *yopH*).
 - 25 - *Autographa californica* nuclear polyhedrosis virus 19 Kd PTPase.
- Dual specificity PTPases.**
- DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185.
 - 30 - DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues.
 - DUSP3 (VHR).
 - DUSP4 (HVH2).

- DUSP5 (HvH3).
- DUSP6 (Pyst1; MKP-3).
- DUSP7 (Pyst2; MKP-X).
- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.
- 5 - Yeast YVH1.
- Vaccinia virus H1 PTPase; a dual specificity phosphatase.

Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length
10 extracellular domain, followed by a transmembrane region and a C-terminal catalytic
cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III)
repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains
in their extracellular region. The cytoplasmic region generally contains two copies of the
PTPase domain. The first seems to have enzymatic activity, while the second is inactive but
15 seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is
generally conserved but some other, presumably important, residues are not.

In the following table, the domain structure of known receptor PTPases is shown:

20		Extracellular	Intracellular				
		-----	-----				
		Ig FN-3	CAH MAM	PTPase			
	Leukocyte common antigen (LCA) (CD45)	0	2	0	0	2	
25	Leukocyte antigen related (LAR)	3	8	0	0	2	
	Drosophila DLAR	3	9	0	0	2	
	Drosophila DPTP	2	2	0	0	2	
	PTP-alpha (LRP)	0	0	0	0	2	
	PTP-beta	0	16	0	0	1	
30	PTP-gamma	0	1	1	0	2	
	PTP-delta	0	>7	0	0	2	
	PTP-epsilon	0	0	0	0	2	
	PTP-kappa	1	4	0	1	2	

					657
PTP-mu	1	4	0	1	2
PTP-zeta	0	1	1	0	2

PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.

A signature pattern was derived for PTPase domains centered on the active site cysteine.

There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.

Consensus pattern [LIVMF][LIVMF SEQ ID NO:2]-H-C-x(2)-G-x(3)-[STC]-[STAGP][STAGP SEQ ID NO:213]-x-[LIVMFY][LIVMFY SEQ ID NO:18] [C is the active site residue]

Notethe M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases.

Notethis documentation entry is linked to both a signature pattern and to profiles. As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.

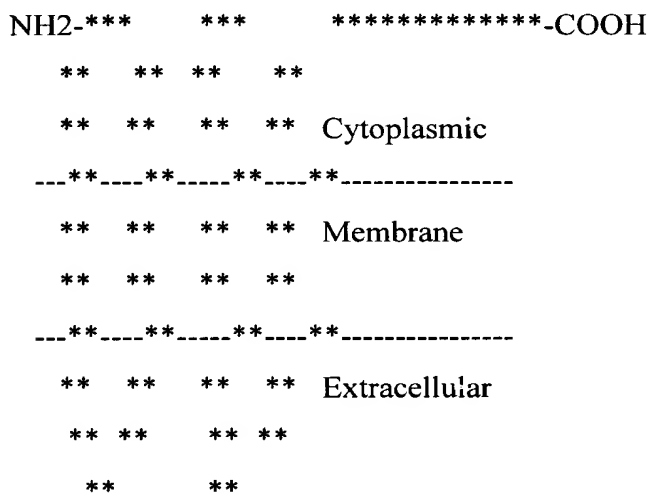
- [1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).
- [2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).
- [3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).
- [4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).
- [5] Hunter T. Cell 58:1013-1016(1989).

780. Connexins signatures

Gap junctions [1] are specialized regions of the plasma membrane which consist of closely packed pairs of transmembrane channels, the connexons, through which small molecules diffuse from a cell to a neighboring cell. Each connexon is composed of an hexamer of an integral membrane protein which is often referred to as connexin. In a given species there are a number of different, yet structurally related, tissue specific, forms of connexins. The types of connexins which are currently known are listed below.

- Connexin 56 (Cx56).
- Connexin 50 (Cx50) (lens fiber protein MP70).
- Connexin 46 (Cx46) (alpha-3).
- Connexin 45 (Cx45) (alpha-6).
- 5 - Connexin 43 (Cx43) (alpha-1).
- Connexin 40 (Cx40) (alpha-5).
- Connexin 38 (Cx38) (alpha-2).
- Connexin 37 (Cx37) (alpha-4).
- Connexin 33 (Cx33) (alpha-7).
- 10 - Connexin 32 (Cx32) (beta-1).
- Connexin 31.1 (Cx31.1) (beta-4).
- Connexin 31 (Cx31) (beta-3).
- Connexin 30.3 (Cx30.3) (beta-5).
- Connexin 26 (Cx26) (beta-2).

15 Structurally the connexins consist of a short cytoplasmic N-terminal domain, followed by four transmembrane segments that delimit two extracellular and one cytoplasmic loops; the C-terminal domain is cytoplasmic and its length is variable (from 20 residues in Cx26 to 260 residues in Cx56). The schematic representation of this structure is shown below.



30 The sequences of the two extracellular loops are well conserved. In both loops there are three conserved cysteines which are involved in disulfide bonds. A signature patterns from each of these two loop regions has been built.

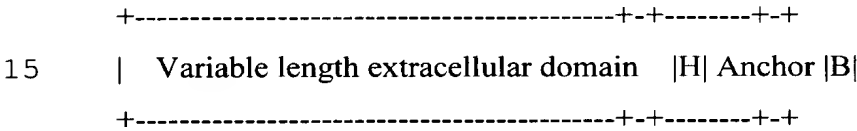
Consensus patternC-[DN]-T-x-Q-P-G-C-x(2)-V-C-[FY]-D [The three C's are involved in disulfide bonds] Consensus patternC-x(3,4)-P-C-x(3)-[LIVM][LIVM SEQ ID NO:4]-[DEN]-C-[FY]-[LIVM][LIVM SEQ ID NO:4]-[SA]-[KR]-P [The three C's are involved in disulfide bonds]

5

[1] Goodenough D.A., Goliger J.A., Paul D.L. Annu. Rev. Biochem. 65:475-502(1996).

781. Gram-positive cocci surface proteins 'anchoring' hexapeptide

10 Surface proteins from Gram-positive cocci contains a conserved hexapeptide located a few residues downstream of a hydrophobic C-terminal membrane anchor region which is followed by a cluster of basic amino acids [1]. This structure is represented in the following schematic representation:



'H': conserved hexapeptide.

'B': cluster of basic residues.

20 It has been proposed that this hexapeptide sequence is responsible for a post-translational modification necessary for the proper anchoring of the proteins which bear it, to the cell wall. Proteins known to contain such hexapeptide are listed below:

- Aggregation substance from streptococcus faecalis (asa1).
- C5a peptidase from Streptococcus pyogenes (scpA).
- 25 - C protein alpha-antigen from Streptococcus agalactiae (bca).
- Cell surface antigen I/II (PAC) from Streptococcus mutans.
- Dextranase from Streptococcus downei (dex).
- Fibronectin-binding protein from Staphylococcus aureus (fnbA).
- Fimbrial subunits from Actinomyces naeslundii and viscosus.
- 30 - IgA binding protein from Streptococcus pyogenes (arp4).
- IgA binding protein (B antigen) from Streptococcus agalactiae (bag).
- IgG binding proteins from Streptococci and Staphylococcus aureus.
- Internalin A from Listeria monocytogenes (inlA).

- M proteins from streptococci.
- Muramidase-released protein from *Streptococcus suis* (mrp).
- Nisin leader peptide processing protease from *Lactococcus lactis* (nisP).
- Protein A from *Staphylococcus aureus*.
- 5 - Trypsin-resistant surface T protein from streptococci.
- Wall-associated protein from *Streptococcus mutans* (wapA).
- Wall-associated serine proteinases from *Lactococcus lactis*.

Consensus pattern L-P-x-T-G-[STGAVDE][STGAVDE SEQ ID NO:696]

10

[1] Schneewind O., Jones K.F., Fischetti V.A. J. Bacteriol. 172:3310-3317(1990).

782. Gamma-glutamyltranspeptidase signature

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the
 15 gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or
 water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway
 for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an
 enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a
 single chain precursor. The active site of GGT is known to be located in the light subunit.

20 The sequences of mammalian and bacterial GGT show a number of regions of high
 similarity [2]. *Pseudomonas cephalosporin acylases* (EC 3.5.1.-) that convert 7-beta-(4-
 carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid
 (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity
 [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

25 One of the conserved regions correspond to the N-terminal extremity of the mature
 light chains of these enzymes. This region has been used as a signature pattern.

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA][LIVMA SEQ ID NO:30]-x(4)-G-[SN]-x-V-
 [STA]-x-T-x-T-[LIVM][LIVM SEQ ID NO:4]-[NE]-x(1,2)-[FY]-G

30

[1] Tate S.S., Meister A. Meth. Enzymol. 113:400-419(1985).

[2] Suzuki H., Kumagai H., Echigo T., Tochikura T. J. Bacteriol. 171:5169-5172(1989).

[3] Ishiye M., Niwa M. Biochim. Biophys. Acta 1132:233-239(1992).

783. Ferrochelatase signature

Ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) [1,2] catalyzes the last step in heme biosynthesis: the chelation of a ferrous ion to proto-porphyrin IX, to form protoheme.

In eukaryotes, ferrochelatase is a mitochondrial protein bound to the inner membrane, whose active site faces the mitochondrial matrix. The mature form of eukaryotic ferrochelatase is composed of about 360 amino acids. In bacteria, ferrochelatase (gene hemH) [3] is a protein of from 310 to 380 amino acids.

The human autosomal dominant disease protoporphyria is due to the reduced activity of ferrochelatase.

The signature pattern for this enzyme is based on a conserved region which contains a histidine residue which could be involved in binding iron.

Consensus pattern [LIVMF][LIVMF SEQ ID NO:2]](2)-x-[ST]-x-H-[GS]-[LIVM][LIVM SEQ ID NO:4]]-P-x(4,5)-[DENQKR][DENQKR SEQ ID NO:697]]-x-G-[DP]-x(1,2)-Y

[1] Labbe-Bois R. J. Biol. Chem. 265:7278-7283(1990).

[2] Brenner D.A., Frasier F. Proc. Natl. Acad. Sci. U.S.A. 88:849-853(1991).

[3] Miyamoto K., Nakahigashi K., Nishimura K., Inokuchi H. J. Mol. Biol. 219:393-398(1991).

784. Cellulose-binding domain, bacterial type

The microbial degradation of cellulose and xylans requires several types of enzyme such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of bacterial cellulases has been shown to consist of about 105 amino acid residues [2]. Enzymes known to contain such a domain are:

- Endoglucanase (gene end1) from *Butyrivibrio fibrisolvens*.
- Endoglucanases A (gene cenA) and B (cenB) from *Cellulomonas fimi*.
- Exoglucanases A (gene cbhA) and B (cbhB) from *Cellulomonas fimi*.

- Endoglucanase E-2 (gene celB) from *Thermomonospora fusca*.
- Endoglucanase A (gene celA) from *Microbispora bispora*.
- Endoglucanases A (gene celA), B (celB) and C (celC) from *Pseudomonas fluorescens*.
- Endoglucanase A (gene celA) from *Streptomyces lividans*.
- 5 - Exocellobiohydrolase (gene cex) from *Cellulomonas fimi*.
- Xylanases A (gene xynA) and B (xynB) from *Pseudomonas fluorescens*.
- Arabinofuranosidase C (EC 3.2.1.55) (xylanase C) (gene xynC) from *Pseudomonas fluorescens*.
- Chitinase 63 (EC 3.2.1.14) from *Streptomyces plicatus*.
- 10 - Chitinase C from *Streptomyces lividans*.

The CBD domain is found either at the N-terminal or at the C-terminal extremity of these enzymes. As it is shown in the following schematic representation, there are two conserved cysteines in this CBD domain - one at each extremity of the domain - which have been shown
 15 [3] to be involved in a disulfide bond. There are also four conserved tryptophan residues which could be involved in the interaction of the CBD with polysaccharides.

```

+-----+
|               |
20  xCxxxxWxxxxxNxxxWxxxxxxxxWxxxxxxxxWNxxxxxGxxxxxxxxxxCx
          *****
  
```

'C': conserved cysteine involved in a disulfide bond. '*': position of the pattern.

Consensus pattern W-N-[STAGR][STAGR SEQ ID NO:698]-[STDN][STDN SEQ ID
 25 NO:699]-[LIVM][LIVM SEQ ID NO:4]-x(2)-[GST]-x-[GST]-x(2)-[LIVMFT][LIVMFT
 SEQ ID NO:282)]-[GA]

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

30 [2] Meinke A., Gilkes N.R., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Protein Seq. Data Anal. 4:349-353(1991).

[3] Gilkes N.R., Claeyssens M., Aebersold R., Henrissat B., Meinke A., Morrison H.D., Kilburn D.G., Warren R.A.J., Miller R.C. Jr. Eur. J. Biochem. 202:367-377(1991).

785. Amidases signature

It has been shown [1,2,3] that several enzymes from various prokaryotic and eukaryotic organisms which are involved in the hydrolysis of amides (amidases) are evolutionary related. These enzymes are listed below.

- Indoleacetamide hydrolase (EC 3.5.1.-), a bacterial plasmid-encoded enzyme that catalyzes the hydrolysis of indole-3-acetamide (IAM) into indole-3-acetate (IAA), the second step in the biosynthesis of auxins from tryptophan.

- Acetamidase from *Emericella nidulans* (gene *amdS*), an enzyme which allows acetamide to be used as a sole carbon or nitrogen source.

- Amidase (EC 3.5.1.4) from *Rhodococcus* sp. N-774 and *Brevibacterium* sp. R312 (gene *amdA*). This enzyme hydrolyzes propionamides efficiently, and also at a lower efficiency, acetamide, acrylamide and indoleacetamide.

- Amidase (EC 3.5.1.4) from *Pseudomonas chlororaphis*.

- 6-aminohexanoate-cyclic-dimer hydrolase (EC 3.5.2.12) (nylon oligomers degrading enzyme E1) (gene *nylA*), a bacterial plasmid encoded enzyme which catalyzes the first step in the degradation of 6-aminohexanoic acid cyclic dimer, a by-product of nylon manufacture [4].

- Glutamyl-tRNA(Gln) amidotransferase subunit A [5].

- Mammalian fatty acid amide hydrolase (gene *FAAH*) [6].

- A putative amidase from yeast (gene *AMD2*).

- *Mycobacterium tuberculosis* putative amidases *amiA2*, *amiB2*, *amiC* and *amiD*.

All these enzymes contain in their central section a highly conserved region rich in glycine, serine, and alanine residues. This region has been used as a signature pattern.

Consensus pattern: G-[GA]-S-[GS]-[GS]-G-x-[GSA]-[GSAVY][GSAVY SEQ ID NO:700)]-x-[LIVM][LIVM SEQ ID NO:4)]-[GSA]-x(6)-[GSAF][GSAF SEQ ID NO:100)]-x-[GA]-x-[DE]-x-[GA]-x-S-[LIVM][LIVM SEQ ID NO:4)]-R-x-P-[GSAC][GSAC SEQ ID NO:93)]

[1] Mayaux J.-F., Cerbelaud E., Soubrier F., Faucher D., Petre D. J. *Bacteriol.* 172:6764-6773(1990).

[2] Hashimoto Y., Nishiyama M., Ikehata O., Horinouchi S., Beppu T. *Biochim. Biophys. Acta* 1088:225-233(1991).

[3] Chang T.-H., Abelson J. *Nucleic Acids Res.* 18:7180-7180(1990).

[4] Tsuchiya K., Fukuyama S., Kanzaki N., Kanagawa K., Negoro S., Okada H. *J. Bacteriol.* 171:3187-3191(1989).

[5] Curnow A.W., Hong K.W., Yuan R., Kim S.I., Martins O., Winkler W., Henkin T.M., Soll D. *Proc. Natl. Acad. Sci. U.S.A.* 94:11819-11826(1997).

[6] Cravatt B.F., Giang D.K., Mayfield S.P., Boger D.L., Lerner R.A., Gilula N.B. *Nature* 384:83-87(1996).

786. Glycosyl hydrolases family 10 active site

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family F [3] or as the glycosyl hydrolases family 10 [4,E1]. The enzymes which are currently known to belong to this family are listed below.

- *Aspergillus awamori* xylanase A (xynA).
- *Bacillus* sp. strain 125 xylanase (xynA).
- *Bacillus stearothermophilus* xylanase.
- *Butyrivibrio fibrisolvens* xylanases A (xynA) and B (xynB).
- *Caldocellum saccharolyticum* bifunctional endoglucanase/exoglucanase (celB). This protein consists of two domains; it is the N-terminal domain, which has exoglucanase activity, which belongs to this family.
- *Caldocellum saccharolyticum* xylanase A (xynA).
- *Caldocellum saccharolyticum* ORF4. This hypothetical protein is encoded in the xynABC operon and is probably a xylanase.
- *Cellulomonas fimi* exoglucanase/xylanase (cex).
- *Clostridium stercorarium* thermostable celloxylanase.
- *Clostridium thermocellum* xylanases Y (xynY) and Z (xynZ).
- *Cryptococcus albidus* xylanase.
- *Penicillium chrysogenum* xylanase (gene xylP).

665

- *Pseudomonas fluorescens* xylanases A (xynA) and B (xynB).

- *Ruminococcus flavefaciens* bifunctional xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn and Trp, and a C-terminal

5 xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases.

- *Streptomyces lividans* xylanase A (xlnA).

- *Thermoanaerobacter saccharolyticum* endoxylanase A (xynA).

- *Thermoascus aurantiacus* xylanase.

- Thermophilic bacterium Rt8.B4 xylanase (xynA).

10 One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the exoglucanase from *Cellulomonas fimi*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region has been used as a signature pattern.

15 Consensus pattern[GTA]-x(2)-[LIVN][LIVN SEQ ID NO:682])-x-[IVMF][IVMF SEQ ID NO:701)]-[ST]-E-[LIY]-[DN]-[LIVMF][LIVMF SEQ ID NO:2)] [E is the active site residue]

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

20 [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

25 [5] Tull D., Withers S.G., Gilkes N.R., Kilburn D.G., Warren R.A.J., Aebersold R. J. Biol. Chem. 266:15621-15625(1991).

787. Fructose-bisphosphate aldolase class-II signatures

Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6- bisphosphate into

30 dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2], mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion – generally zinc - for their activity.

This family also includes the following proteins:

- Escherichia coli galactitol operon protein gatY which catalyzes the transformation of tagatose 1,6-bisphosphate into glycerone phosphate and D- glyceraldehyde 3-phosphate.
- 5 - Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY.

As signature patterns for this class of enzyme, two conserved regions were selected. The first pattern is located in the first half of the sequence and contains two histidine residues that have
 10 been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.

Consensus pattern[FYVMT][FYVMT SEQ ID NO:702)]-x(1,3)-[LIVMH][LIVMH SEQ ID NO:703)]-[APN]-[LIVM][LIVM SEQ ID NO:4)]-x(1,2)-[LIVM][LIVM SEQ ID NO:4)]-H-
 15 x-D-H- [GACH][GACH SEQ ID NO:704)] [The two H's are zinc ligands]
 Consensus pattern[LIVM][LIVM SEQ ID NO:4)]-E-x-E-[LIVM][LIVM SEQ ID NO:4)]-G-x(2)-[GM]-[GSTA][GSTA SEQ ID NO:19)]-x-E

- [1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).
- 20 [2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).
- [3] von der Osten C.H., Barbas C.F. III, Wong C.-H., Sinskey A.J. Mol. Microbiol. 3:1625-1637(1989).
- [4] Berry A., Marshall K.E. FEBS Lett. 318:11-16(1993).

25 788. Prolyl oligopeptidase family serine active site

The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

- 30 - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium

meningosepticum and *Aeromonas hydrophila*); there is a high degree of sequence conservation between these sequences.

- *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.

5 - Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

10 - Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

15 A conserved serine residue has experimentally been shown (in *E. coli* protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

20

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW][LIVMFYW SEQ ID NO:26]-x(14)-G-x-S-x-G-G-[LIVMFYW][LIVMFYW SEQ ID NO:26]](2) [S is the active site residue]

Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases

25 [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).

[3] Polgar L., Szabo E.

30 Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

Formate--tetrahydrofolate ligase (EC 6.3.4.3) (formyltetrahydrofolate synthetase) (FTHFS) is one of the enzymes participating in the transfer of one-carbon units, an essential element of various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by FTHFS, methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).

In eukaryotes the FTHFS activity is expressed by a multifunctional enzyme, C-1-tetrahydrofolate synthase (C1-THF synthase), which also catalyzes the dehydrogenase and cyclohydrolase activities. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the FTHFS domain consist of about 600 amino acid residues and is located in the C-terminal section of C1-THF synthase. In prokaryotes FTHFS activity is expressed by a monofunctional homotetrameric enzyme of about 560 amino acid residues [2].

The sequence of FTHFS is highly conserved in all forms of the enzyme. As signature patterns, two regions that are almost perfectly conserved were selected. The first one is a glycine-rich segment located in the N-terminal part of FTHFS and which could be part of an ATP-binding domain [2]. The second pattern is located in the central section of FTHFS.

Consensus pattern G-[LIVM][LIVM SEQ ID NO:4]-K-G-G-A-A-G-G-G-Y
Consensus pattern V-A-T-[IV]-R-A-L-K-x-[HN]-G-G

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).

[2] Lovell C.R., Przybyla A., Ljungdahl L.G. Biochemistry 29:5687-5694(1990).

790. Transthyretin signatures

Transthyretin (prealbumin) [1] is a thyroid hormone-binding protein that seems to transport thyroxine (T4) from the bloodstream to the brain. It is a protein of about 130 amino acids that assembles as a homotetramer and forms an internal channel that binds thyroxine. Transthyretin is mainly synthesized in the brain choroid plexus. In humans, variants of the protein are associated with distinct forms of amyloidosis.

The sequence of transthyretin is highly conserved in vertebrates. A number of uncharacterized proteins also belong to this family:

- Escherichia coli hypothetical protein yedX.
- Bacillus subtilis hypothetical protein yunM.
- Caenorhabditis elegans hypothetical protein R09H10.3.
- Caenorhabditis elegans hypothetical protein ZK697.8.

5

Two regions were selected as signature patterns. The first located in the N-terminal extremity starts with a lysine known to be involved in binding T4. The second pattern is located in the C-terminal extremity.

10 Consensus pattern[KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine]
Consensus patternY-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS]

[1] Schreiber G., Richardson S.J. Comp. Biochem. Physiol. 116B:137-160(1997).

15 791. Dihydropteroate synthase signatures

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes that are involved in the biosynthesis of folates are therefore the target of a variety of antimicrobial agents such as trimethoprim or sulfonamides.

20

Dihydropteroate synthase (EC 2.5.1.15) (DHPS) catalyzes the condensation of 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate to para-aminobenzoic acid to form 7,8-dihydropteroate. This is the second step in the three steps pathway leading from 6-hydroxymethyl-7,8-dihydropteridine to 7,8-dihydrofolate. DHPS is the target of sulfonamides which are substrates analog that compete with para-aminobenzoic acid.

25

Bacterial DHPS (gene sul or folP) [1] is a protein of about 275 to 315 amino acid residues which is either chromosomally encoded or found on various antibiotic resistance plasmids. In the lower eukaryote Pneumocystis carinii, DHPS is the C-terminal domain of a multifunctional folate synthesis enzyme (gene fas) [2].

30

Two signature patterns for DHPS were developed, the first signature is located in the N-terminal section of these enzymes, while the second signature is located in the central section.

670

Consensus pattern[LIVM][LIVM SEQ ID NO:4]-x-[AG]-[LIVMF][LIVMF SEQ ID NO:2](2)-N-x-T-x-D-S-F-x-D-x-[SG]

Consensus pattern[GE]-[SA]-x-[LIVM][LIVM SEQ ID NO:4](2)-D-[LIVM][LIVM SEQ ID NO:4]-G-[GP]-x(2)-[STA]-x-P

5

[1] Slock J., Stahly D.P., Han C.-Y., Six E.W., Crawford I.P. J. Bacteriol. 172:7211-7226(1990).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

10

792. Phosphatidylinositol 3- and 4-kinases signatures

Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known:

15

- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain (p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at least two different types of p100 subunits (alpha and beta).

20

- Yeast TOR1/DDR1 and TOR2/DDR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd.

- Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd.

- Arabidopsis thaliana and soybean VPS34 homologs.

25

Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are known:

30

- Human PI4-kinase alpha.
- Yeast PIK1, a nuclear protein of 120 Kd.
- Yeast STT4, a protein of 214 Kd.

The PI3- and PI4-kinases share a well conserved domain at their C-terminal section; this domain seems to be distantly related to the catalytic domain of protein kinases [2]. Two signature patterns were developed from the best conserved parts of this domain.

5 Four additional proteins belong to this family:

- Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex.

- Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination.

10 - Yeast protein TEL1 which is involved in controlling telomere length.

- Yeast hypothetical protein YHR099w, a distantly related member of this family.

- Fission yeast hypothetical protein SpAC22E12.16C.

Consensus pattern[LIVMFAC][LIVMFAC SEQ ID NO:95]-K-x(1,3)-[DEA]-[DE]-

15 [LIVMC][LIVMC SEQ ID NO:142]-R-Q-[DE]-x(4)-Q

Consensus pattern[GS]-x-[AV]-x(3)-[LIVM][LIVM SEQ ID NO:4]-x(2)-[FYH]-

[LIVM][LIVM SEQ ID NO:4]](2)-x-[LIVMF][LIVMF SEQ ID NO:2]-x-D-R-H-x(2)-N

[1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea
20 F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell
70:419-429(1992).

[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell
73:585-596(1993).

[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science
25 260:88-91(1993).

[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J. 13:2352-
2361(1994).

[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L.
Nature 369:756-758(1994).

30 [6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).

793. FAD-dependent glycerol-3-phosphate dehydrogenase signatures

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In bacteria [1] it is associated with the utilization of glycerol coupled to respiration. In *Escherichia coli*, two isozymes are known: one expressed under anaerobic conditions (gene *glpA*) and one in aerobic conditions (gene *glpD*). In eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,3].

These enzymes are proteins of about 60 to 70 Kd which contain a probable FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs from the bacterial or yeast proteins by having an EF-hand calcium-binding region (See <PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-binding domain and one which corresponds to a conserved region in the central part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV][STACV SEQ ID NO:146]-G-x-A-x-D-x(3)-R-G

Consensus patternG-G-K-x(2)-[GSTE][GSTE SEQ ID NO:705]-Y-R-x(2)-A

[1] Austin D., Larson T.J. J. Bacteriol. 173:101-107(1991).

[2] Roennow B., Kielland-Brandt M.C. Yeast 9:1121-1130(1993).

[3] Brown L.J., McDonald M.J., Lehn D.A., Moran S.M. J. Biol. Chem. 269:14363-14366(1994).

794. NOL1/NOP2/sun family signature

The following proteins seems to be evolutionary related:

- Mammalian proliferating-cell nucleolar antigen p120 (gene NOL1) which may play a role in the regulation of the cell cycle and the increased nucleolar activity that is associated with the cell proliferation.

- Yeast nucleolar protein NOP2 (or YNA1) which could be involved in nucleolar function during the onset of growth, and in the maintenance of nucleolar structure.

- Yeast hypothetical protein YBL024w.

- Bacterial protein sun (also known as *fm*).

673

- Escherichia coli hypothetical protein yebU.
- Mycobacterium tuberculosis hypothetical protein MtCY21B4.24.
- Methanococcus jannaschii hypothetical protein MJ0026.

5 NOL1 is a protein of 855 residues, NOP2 consists of 618 residues, YBL024w of 684, sun is a protein of about 430 to 450 residues and MJ026 has 274 residues. They share a conserved central domain which contains some highly conserved regions. One of these regions was selected as a signature pattern.

10 Consensus pattern[FV]-D-[KRA]-[LIVMA][LIVMA SEQ ID NO:30]-L-x-D-[AV]-P-C-[ST]-[GA]

795. moaA / nifB / pqqE family signature

15 A number of proteins involved in the biosynthesis of metallo cofactors have been shown [1,2] to be evolutionary related. These proteins are:

- Bacterial and archeobacterial protein moaA, which is involved in the biosynthesis of the molybdenum cofactor (molybdopterin; MPT).
- Arabidopsis thaliana cnx2, a protein involved in molybdopterin biosynthesis and which is highly similar to moaA.
- 20 - Bacillus subtilis narA, which seems to be the moaA ortholog in that bacteria.
- Bacterial protein nifB (or fixZ) which is involved in the biosynthesis of the nitrogenase iron-molybdenum cofactor.
- Bacterial protein pqqE which is involved in the biosynthesis of the cofactor pyrrolo-quinoline-quinone (PQQ).
- 25 - Pyrococcus furiosus cmo, a protein involved in the synthesis of a molybdopterin-based tungsten cofactor.
- Caenorhabditis elegans hypothetical protein F49E2.1.

30 All these proteins share, in their N-terminal region, a conserved domain that contains three cysteines. In moaA, these cysteines have been shown [1] to be important for the biological activity. They could be involved in the binding of an iron-sulfur cluster.

Consensus pattern[LIV]-x(3)-C-[NP]-[LIVMF][LIVMF SEQ ID NO:2]-[QRS]-C-x-[FYM]-
C [The three C's are putative Fe-S ligands]

[1] Menendez C., Igloi G., Henninger H., Brandsch R. Arch. Microbiol. 164:142-151(1995).

5 [2] Hoff T., Schnorr K.M., Meyer C., Caboche M. J. Biol. Chem. 270:6100-6107(1995).

796. Forkhead-associated (FHA) domain profile

The forkhead-associated (FHA) domain [1,E1] is a putative nuclear signalling domain found in a variety of otherwise unrelated proteins. The FHA domain comprise approximately
10 55 to 75 amino acids and contains three highly conserved blocks separated by divergent spacer regions. Currently it has been found in the following proteins:

- Four transcription factors that also contain a forkhead (FH) domain: mouse myocyte nuclear factor 1 (MNF1), yeast transcription factor FHL1, which probably controls pre-mRNA processing, and yeast FKH1 and FKH2. In those protein the FHA domain is located
15 N-terminal of the DNA-binding FH domain.

- Kinase-associated protein phosphatase (KAPP) from *Arabidopsis thaliana*, a protein which specifically interacts with the receptor-type Ser/Thr-kinase RLK5. In KAPP, the FHA domain maps to a region that interacts with the receptor-type protein kinase RLK5 only if the kinase is phosphorylated on serine residues [2].

20 - Two protein kinases from yeast that are involved in mediating the nuclear response to DNA damage: DUN1 and SPK1/SAD1 [3]. The latter is the only known protein containing two copies of the FHA domain.

- Protein kinase cds1 from fission yeast contains a FHA domain and might be the ortholog of SPK1.

25 - Protein kinase MEK1 from yeast, which is involved in meiotic recombination.

- Human nuclear antigen Ki67 which is expressed only in proliferating cells.

- Yeast hypothetical protein YHR115c, which contains a RING-finger C-terminal of the FHA domain.

- Yeast hypothetical proteins L8083.1 and 9346.10, which contain an extensive coiled-coil
30 region C-terminal of the FHA domain.

- *Caenorhabditis elegans* hypothetical protein ZK632.2.

- *Caenorhabditis elegans* hypothetical protein C01G6.5.

675

- FraH from the prokaryote *Anabaena*, which contains a zinc-finger motif N-terminal of the FHA domain.

- An ORF from the bacterium *Streptomyces*, which is on the opposite strand of the protein kinase *pksl*, overlapping the ORF of the kinase.

5

[1] Hofmann K.O., Bucher P. Trends Biochem. Sci. 20:347-349(1995).

[2] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. Science 266:793-795(1994).

[3] Navas T.A., Zhou Z., Elledge S.J. Cell 80:29-39(1995).

10

797. Ald_Xan_dh_C

Aldehyde oxidase and xanthine dehydrogenase, C terminus

[1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*." Science 1995;270:1170-1176.

15

Number of members: 54

20

798. Glyco_hydro_38

Glycosyl hydrolases family 38

Glycosyl hydrolases are key enzymes of carbohydrate metabolism.

Number of members: 20

25

[1] Henrissat B; Medline: 98313424; "Glycosidase families" Biochem Soc Trans 1998;26:153-156.

799. HECT

30

HECT-domain (ubiquitin-transferase).

The name HECT comes from Homologous to the E6-AP Carboxyl Terminus.

Number of members: 43

[1] Huibregtse JM, Scheffner M, Beaudenon S, Howley PM; Medline: 95223981; "A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase." Proc Natl Acad Sci U S A 1995;92:2563-2567.

800. HRDC

HRDC domain

The HRDC (Helicase and RNase D C-terminal) domain has a putative role in nucleic acid binding. Mutations in the HRDC domain cause human disease.

Number of members: 19

[1] Morozov V, Mushegian AR, Koonin EV, Bork P; Medline: 98060076; "A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases" Trends Biochem Sci 1997;22:417-418.

801. Integrase

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain. The central domain is the catalytic domain [1]. The carboxyl terminal domain is a DNA binding domain [2].

Number of members: 581

[1] Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Medline: 95099322. "Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases." Science 1994;266:1981-1986.

[2] Lodi PJ, Ernst JA, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM, Gronenborn AM; Medline: 95359147; "Solution structure of the DNA binding domain of HIV-1 integrase." Biochemistry 1995;34:9826-9833

802. lig_chan

Ligand-gated ion channel

This family includes the four transmembrane regions of the ionotropic glutamate receptors and NMDA receptors.

5 Number of members: 128

[1] Tong G, Shepherd D, Jahr CE; Medline: 95184014; "Synaptic desensitization of NMDA receptors by calcineurin." Science 1995;267:1510-1512.

10 803. RhoGAP

RhoGAP domain

GTPase activator proteins towards Rho/Rac/Cdc42-like small GTPases.

Number of members: 97

15

[1] Musacchio A, Cantley LC, Harrison SC; Medline: 97121392; "Crystal structure of the breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 alpha subunit." Proc Natl Acad Sci U S A 1996;93:14373-14378.

20 [2] Barrett T, Xiao B, Dodson EJ, Dodson G, Ludbrook SB, Nurmahomed K, Gamblin SJ, Musacchio A, Smerdon SJ, Eccleston JF; Medline: 97162209; "The structure of the GTPase-activating domain from p50rhoGAP." Nature 1997;385:458-461.

[3] Rittinger K, Walker PA, Eccleston JF, Nurmahomed K, Owen D, Laue E, Gamblin SJ, Smerdon SJ; Medline: 97404320; "Crystal structure of a small G protein in complex with the GTPase-activating protein rhoGAP." Nature 1997;388:693-697.

25 [4] Boguski MS, McCormick F; Medline: 94081948; "Proteins regulating Ras and its relatives." Nature 1993;366:643-654.

804. vwd

von Willebrand factor type D domain

30

[1] Bork P; Medline: 93327926; "The modular architecture of a new family of growth regulators related to connective tissue growth factor." FEBS lett 1993;327:125-130.

Number of members: 92

805. zf-C4_Topoiso

Topoisomerase DNA binding C4 zinc finger

5

[1] Tse-Dinh YC, Beran-Steed RK; Medline: 89034032; "Escherichia coli DNA topoisomerase I is a zinc metalloprotein with three repetitive zinc-binding domains." J Biol Chem 1988;263:15857-15859.

10

[2] Ahumada A, Tse-Dinh YC; Medline: 99011409; "The Zn(II) binding motifs of E. coli DNA topoisomerase I is part of a high-affinity DNA binding domain." Biochem Biophys Res Commun 1998;251:509-514.

Number of members: 51

15

806. AIRC

AIR carboxylase

Members of this family catalyse the decarboxylation of 1-(5-phosphoribosyl)-5-amino-4-imidazole-carboxylate (AIR). This family catalyse the sixth step of de novo purine

20

biosynthesis. Some members of this family contain two copies of this domain. Number of members: 35

807. Bromodomain signature and profile

PROSITE cross-reference(s): PS00633; BROMODOMAIN_1, PS50014;

25

BROMODOMAIN_2

The bromodomain [1,2,3] is a conserved region of about 70 amino acids found in the following proteins:

30

- Higher eukaryotes transcription initiation factor TFIID 250 Kd subunit (TBP-associated factor p250) (gene CCG1). P250 associated with the TFIID TATA-box binding protein and seems essential for progression of the G1 phase of the cell cycle.

- Human RING3, a protein of unknown function encoded in the MHC class II locus.

- Mammalian CREB-binding protein (CBP), which mediates cAMP-gene regulation by binding specifically to phosphorylated CREB protein.

- Drosophila female sterile homeotic protein (gene fsh), required maternally for proper expression of other homeotic genes involved in pattern formation, such as Ubx.

5 - Drosophila brahma protein (gene brm), a protein required for the activation of multiple homeotic genes.

- Mammalian homologs of brahma. In human, three brahma-like proteins are known: SNF2a(hBRM), SNF2b, and BRG1.

- Human BS69, a protein that binds to adenovirus E1A and inhibits E1A transactivation

10 - Human peregrin (or Br140).

- Yeast BDF1 [3], a transcription factor involved in the expression of a broad class of genes including snRNAs.

- Yeast GCN5, a general transcriptional activator operating in concert with certain other DNA-binding transcriptional activators, such as GCN4, HAP2/3/4 or ADA2.

15 - Yeast NPS1/STH1, involved in G(2) phase control in mitosis.

- Yeast SNF2/SWI2, which is part of a complex with the SNF5, SNF6, SWI3 and ADR6/SWI1 proteins. This SWI-complex is involved in transcriptional activation.

- Yeast SPT7, a transcriptional activator of Ty elements and possibly other genes.

- Caenorhabditis elegans protein cbp-1.

20 - Yeast hypothetical protein YGR056w.

- Yeast hypothetical protein YKR008w.

- Yeast hypothetical protein L9638.1.

Some proteins contain a region which, while similar to some extent to a classical
25 bromodomain, diverges from it by either lacking part of the domain or because of an insertion. These proteins are:

- Mammalian protein HRX (also known as All-1 or MLL), a protein involved in translocations leading to acute leukemias and which possibly acts as a transcriptional
30 regulatory factor. HRX contains a region similar to the C- terminal half of the bromodomain.

- Caenorhabditis elegans hypothetical protein ZK783.4. The bromodomain of this protein has a 23 amino-acid insertion.

680

- Yeast protein YTA7. This protein contains a region with significant similarity to the C-terminal half of the bromodomain. As it is a member of the AAA family (see <PDOC00572>) it is also in a functionally different context.

- 5 The above proteins generally contain a single bromodomain, but some of them contain two copies, this is the case of BDF1, CCG1, fsh, RING3, YKR008w and L9638.1.

The exact function of this domain is not yet known but it is thought to be involved in protein-protein interactions and it may be important for the assembly or activity of multicomponent
10 complexes involved in transcriptional activation.

The consensus pattern that has been developed spans a major part of the bromodomain; a more sensitive detection is available through the use of a profile which spans the whole domain.

15

Consensus pattern[STANVF][STANVF SEQ ID NO:706]]-x(2)-F-x(4)-[DNS]-x(5,7)-
[DENQTF][DENQTF SEQ ID NO:707]]-Y-[HFY]-x(2)-
[LIVMFY][LIVMFY SEQ ID NO:18]]-x(3)-[LIVM][LIVM SEQ ID NO:4]]-x(4)-
[LIVM][LIVM SEQ ID NO:4]]-x(6,8)-Y-x(12,13)-[LIVM][LIVM SEQ ID NO:4]]-
20 x(2)-N-[SACF][SACF SEQ ID NO:708]]-x(2)-[FY]

References

- [1] Haynes S.R., Doolard C., Winston F., Beck S., Trowsdale J., Dawid I.B. Nucleic Acids Res. 20:2693-2603(1992).
- 25 [2] Tamkun J.W., Deuring R., Scott M.P., Kissinger M., Pattatucci A.M., Kaufman T.C., Kennison J.A. Cell 68:561-572(1992).
- [3] Tamkun J.W. Curr. Opin. Genet. Dev. 5:473-477(1995).

808. (CH) Actinin-type actin-binding domain signatures

30 PROSITE cross-reference(s): PS00019; ACTININ_1, PS00020; ACTININ_2

Alpha-actinin is a F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures [1]. The actin-binding domain of alpha-actinin seems to reside in the

first 250 residues of the protein. A similar actin-binding domain has been found in the N-terminal region of many different actin-binding proteins [2,3]:

- In the beta chain of spectrin (or fodrin).
- 5 - In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane.
- In the slime mold gelation factor (or ABP-120).
- In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins.
- 10 - In fimbrin (or plastin), an actin-bundling protein. Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein.

Two conserved regions were selected as signature patterns for this type of main. The first of this region is located at the beginning of the domain, hile the second one is located in the central section and has been shown to be essential for the binding of actin.

Consensus pattern[EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N

Consensus pattern[LIVM][LIVM SEQ ID NO:4)]-x-[SGN]-[LIVM][LIVM SEQ ID NO:4)]-
 20 [DAGHE][DAGHE SEQ ID NO:709)]-[SAG]-x-[DNEAG][DNEAG SEQ ID NO:710)]-
 [LIVM][LIVM SEQ ID NO:4)]-x-
 [DEAG][DEAG SEQ ID NO:711)]-x(4)-[LIVM][LIVM SEQ ID NO:4)]-x-[LM]-[SAG]-
 [LIVM][LIVM SEQ ID NO:4)]-[LIVMT][LIVMT SEQ ID NO:1)]-W-x- [LIVM][LIVM
 SEQ ID NO:4)](2)

25

- [1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).
- [2] Matsudaira P. Trends Biochem. Sci. 16:87-92(1991).
- [3] Dubreuil R.R. BioEssays 13:219-226(1991).

30 809. (COX1) Heme-copper oxidase subunit I, copper B binding region signature
 PROSITE cross-reference(s): PS00077; COX1
 Heme-copper respiratory oxidases [1] are oligomeric integral membrane protein
 complexes that catalyze the terminal step in the respiratory chain: they

transfer electrons from cytochrome c or a quinol to oxygen. Some terminal oxidases generate a transmembrane proton gradient across the plasma membrane (prokaryotes) or the mitochondrial inner membrane (eukaryotes). The enzyme complex consists of 3-4 subunits (prokaryotes) up to 13 polypeptides (mammals) of which only the catalytic subunit (equivalent to mammalian subunit 1 (CO I)) is found in all heme-copper respiratory oxidases. The presence of a bimetallic center (formed by a high-spin heme and copper B) as well as a low-spin heme, both ligated to six conserved histidine residues near the outer side of four transmembrane spans within CO I is common to all family members [2-4].

In contrary to eukaryotes the respiratory chain of prokaryotes is branched to multiple terminal oxidases. The enzyme complexes vary in heme and copper composition, substrate type and substrate affinity. The different respiratory oxidases allow the cells to customize their respiratory systems according a variety of environmental growth conditions [1].

Recently also a component of an anaerobic respiratory chain has been found to contain the copper B binding signature of this family: nitric oxide reductase (NOR) exists in denitrifying species of Archae and Eubacteria.

Enzymes that belong to this family are:

- Mitochondrial-type cytochrome c oxidase (EC 1.9.3.1) which uses cytochrome c as electron donor. The electrons are transferred via copper A (Cu(A)) and heme a to the bimetallic center of CO I that is formed by a penta-coordinated heme a and copper B (Cu(B)). Subunit 1 contains 12 transmembrane regions. Cu(B) is said to be ligated to three of the conserved histidine residues within the transmembrane segments 6 and 7.
- Quinol oxidase from prokaryotes that transfers electrons from a quinol to the binuclear center of polypeptide I. This category of enzymes includes Escherichia coli cytochrome O terminal oxidase complex which is a component of the aerobic respiratory chain that predominates when cells are grown at high aeration.

- FixN, the catalytic subunit of a cytochrome c oxidase expressed in nitrogen-fixing bacteroids living in root nodules. The high affinity for oxygen allows oxidative phosphorylation under low oxygen concentrations. A similar enzyme has been found in other purple bacteria.

5 - Nitric oxide reductase (EC 1.7.99.7) from *Pseudomonas stutzeri*. NOR reduces nitrate to dinitrogen. It is a heterodimer of norC and the catalytic subunit norB. The latter contains the 6 invariant histidine residues and 12 transmembrane segments [5].

10 As a signature pattern the copper-binding region was used.

Consensus pattern[YWG]-[LIVFYWTA][LIVFYWTA SEQ ID NO:712]](2)-[VGS]-H-[LNP]-x-V-x(44,47)-H-H [The three H's are copper B ligands]

15

Notecytochrome bd complexes do not belong to this family.

[1]

Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B.

20 J. Bacteriol. 176:5587-5600(1994).

[2]

Castresana J., Luebben M., Saraste M., Higgins D.G.

EMBO J. 13:2516-2525(1994).

[3]

25 Capaldi R.A., Malatesta F., Darley-USmar V.M.

Biochim. Biophys. Acta 726:135-148(1983).

[4]

Holm L., Saraste M., Wikstrom M.

EMBO J. 6:2819-2823(1987).

30

[5]

Saraste M., Castresana J.

FEBS Lett. 341:1-4(1994).

810. (dehydrog_molyb) Eukaryotic molybdopterin oxidoreductases signature
PROSITE cross-reference(s): PS00559; MOLYBDOPTERIN_EUK

A number of different eukaryotic oxidoreductases that require and bind a
molybdopterin cofactor have been shown [1] to share a few regions of sequence
similarity. These enzymes are:

- Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of
xanthine to uric acid with the concomitant reduction of NAD. Structurally,
this enzyme of about 1300 amino acids consists of at least three distinct
domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain
(see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Mo-
pterin domain.

- Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation aldehydes into
acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its
sequence and domain structure.

- Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate
to nitrite. Structurally, this enzyme of about 900 amino acids consists of
an N-terminal Mo-pterin domain, a central cytochrome b5-type heme-binding
domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome
reductase domain.

- Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to
sulfate. Structurally, this enzyme of about 460 amino acids consists of an
N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain.

There are a few conserved regions in the sequence of the molybdopterin-binding
domain of these enzymes. The pattern used to detect these proteins is based
on one of them. It contains a cysteine residue which could be involved in
binding the molybdopterin cofactor.

Consensus pattern[GA]-x(3)-[KRNQHT][KRNQHT SEQ ID NO:396]-x(11,14)-
[LIVMFYWS][LIVMFYWS SEQ ID NO:301]-x(8)-[LIVMF][LIVMF SEQ ID NO:2]-x-C-
x(2)-[DEN]-R-x(2)-[DE]

[1]

Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle
W.A., Bray R.C.

5 Biochim. Biophys. Acta 1057:157-185(1991).

811. (DNA_ligase) ATP-dependent DNA ligase signatures

PROSITE cross-reference(s): PS00697; DNA_LIGASE_A1, PS00333; DNA_LIGASE_A2

10 DNA ligase (polydeoxyribonucleotide synthase) is the enzyme that joins two DNA
fragments by catalyzing the formation of an internucleotide ester bond between
phosphate and deoxyribose. It is active during DNA replication, DNA repair and
DNA recombination. There are two forms of DNA ligase: one requires ATP
(EC 6.5.1.1), the other NAD (EC 6.5.1.2).

15

Eukaryotic, archaeobacterial, virus and phage DNA ligases are ATP-dependent.
During the first step of the joining reaction, the ligase interacts with ATP
to form a covalent enzyme-adenylate intermediate. A conserved lysine residue
is the site of adenylation [1,2].

20

Apart from the active site region, the only conserved region common to all
ATP-dependent DNA ligases is found [3] in the C-terminal section and contains
a conserved glutamate as well as four positions with conserved basic residues.

25 Signature patterns were developed for both conserved regions.

Consensus pattern[EDQH][EDQH SEQ ID NO:713]]-x-K-x-[DN]-G-x-R-
[GACIVM][GACIVM SEQ ID NO:714]] [K is the active site
residue]

30

Consensus patternE-G-[LIVMA][LIVMA SEQ ID NO:30]]-[LIVM][LIVM SEQ ID
NO:4]](2)-[KR]-x(5,8)-[YW]-[QNEK][QNEK SEQ ID NO:715]]-x(2,6)-
[KRH]-x(3,5)-K-[LIVMFY][LIVMFY SEQ ID NO:18]]-K

Sequences known to belong to this class detected by the patternALL, except for archebacterial DNA ligases.

[1]

5 Tomkinson A.E., Totty N.F., Ginsburg M., Lindahl T.
Proc. Natl. Acad. Sci. U.S.A. 88:400-404(1991).

[2]

Lindahl T., Barnes D.E.
Annu. Rev. Biochem. 61:251-281(1992).

10 [3]

Kletzin A.
Nucleic Acids Res. 20:5389-5396(1992).

812. (FAD_Gly3P_dh) FAD-dependent glycerol-3-phosphate dehydrogenase signatures
15 PROSITE cross-reference(s): PS00977; FAD_G3PDH_1, PS00978; FAD_G3PDH_2

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In bacteria [1] it is associated with the utilization of glycerol coupled to
20 respiration. In Escherichia coli, two isozymes are known: one expressed under anaerobic conditions (gene glpA) and one in aerobic conditions (gene glpD). In eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2, 3].

25

These enzymes are proteins of about 60 to 70 Kd which contain a probable FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs from the bacterial or yeast proteins by having an EF-hand calcium-binding region (See <PDOC00018>) in its C-terminal extremity.

30

Two signature patterns were developed. One based on the first half of the FAD-binding domain and one which corresponds to a conserved region in the central part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV][STACV SEQ ID NO:146]-G-x-A-x-D-x(3)-R-G

5 Consensus patternG-G-K-x(2)-[GSTE][GSTE SEQ ID NO:705]-Y-R-x(2)-A

[1]

Austin D., Larson T.J.

J. Bacteriol. 173:101-107(1991).

[2]

10 Roennow B., Kielland-Brandt M.C.

Yeast 9:1121-1130(1993).

[3]

Brown L.J., McDonald M.J., Lehn D.A., Moran S.M.

J. Biol. Chem. 269:14363-14366(1994).

15

813. (Fapy_DNA_glyco) Formamidopyrimidine-DNA glycosylase signature
PROSITE cross-reference(s): PS01242; FPG

Formamidopyrimidine-DNA glycosylase (EC 3.2.2.23) [1] (Fapy-DNA glycosylase)

20 (gene fpg) is a bacterial enzyme involved in DNA repair and which excise
oxidized purine bases to release 2,6-diamino-4-hydroxy-5N-methylformamido-
pyrimidine (Fapy) and 7,8-dihydro-8-oxoguanine (8-OxoG) residues. In addition
to its glycosylase activity, FPG can also nick DNA at apurinic/apyrimidinic
sites (AP sites). FPG is a monomeric protein of about 32 Kd which binds and
25 require zinc for its activity.

The binding site for zinc seems to be located in the C-terminal part of the
enzyme where four conserved and essential [2] cysteines are located. A signature pattern
was developed based on this region.

30

Consensus patternC-x(2,4)-C-x-[GTAQ][GTAQ SEQ ID NO:716]-x-[IV]-x(7)-R-
[GSTAN][GSTAN SEQ ID NO:296]-[STA]-x-[FYI]-C- x(2)-C-Q

[The four C's are putative zinc ligands]

[1]

Duwat P., de Oliveira R., Ehrlich S.D., Boiteux S.
Microbiology 141:411-417(1995).

5 [2]

O'Connor T.E., Graves R.J., Demurcia G., Castaing B., Laval J.
J. Biol. Chem. 268:9063-9070(1993).

814. (G_glu_transpept) Gamma-glutamyltranspeptidase signature

10 PROSITE cross-reference(s): PS00462; G_GLU_TRANSPEPTIDASE

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

20

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. Pseudomonas cephalosporin acylases (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

25

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region was used as a signature pattern.

30

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA][LIVMA SEQ ID NO:30]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-

[LIVM][LIVM SEQ ID NO:4)]-[NE]-x(1,2)-[FY]-G

[1]

Tate S.S., Meister A.

5 Meth. Enzymol. 113:400-419(1985).

[2]

Suzuki H., Kumagai H., Echigo T., Tochikura T.

J. Bacteriol. 171:5169-5172(1989).

[3]

10 Ishiye M., Niwa M.

Biochim. Biophys. Acta 1132:233-239(1992).

815. G-protein gamma subunit profile

PROSITE cross-reference(s): PS50058; G_PROTEIN_GAMMA

15

Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

20

The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least 12 different isoforms of gamma subunits.

25

The *Caenorhabditis elegans* protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.

30

A profile was developed that spans the complete length of the gamma subunit.

[1]

Pennington S.R.

Protein Prof. 2:16-315(1995).

5 816. GNS1/SUR4 family signature

PROSITE cross-reference(s): PS01188; GNS1_SUR4

The following group of eukaryotic integral membrane proteins, whose exact function has not yet clearly been established, are evolutionary related [1]:

10

- Yeast GNS1 [2], a protein involved in synthesis of 1,3-beta-glucan.

- Yeast SUR4 (or APA1, SRE1) [3], a protein that could act in a glucose-signaling pathway that controls the expression of several genes that are transcriptionally regulated by glucose.

15

- Yeast hypothetical protein YJL196c.

- Caenorhabditis elegans hypothetical protein C40H1.4.

- Caenorhabditis elegans hypothetical protein D2024.3.

20

The proteins have from 290 to 435 amino acid residues. Structurally, they seem to be formed of three sections: a N-terminal region with two transmembrane domains, a central hydrophilic loop and a C-terminal region that contains from one to three transmembrane domains. A conserved region that contains three histidines was selected as a signature pattern. This region is located in the hydrophilic loop.

25

Consensus pattern L-x-F-L-H-x-Y-H-H

[1]

Bairoch A.

30

Unpublished observations (1996).

[2]

El-Sherbeini M., Clemas J.A.

J. Bacteriol. 177:3227-3234(1995).

[3]

Garcia-Arranz M., Maldonado A.M., Mazon M.J., Portillo F.
J. Biol. Chem. 269:18076-18082(1994).

- 5 817. Immunoglobulins and major histocompatibility complex proteins signature
PROSITE cross-reference(s): PS00290; IG_MHC

The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two
light chains and two heavy chains linked by disulfide bonds. There are two
10 types of light chains: kappa and lambda, each composed of a constant domain
(CL) and a variable domain (VL). There are five types of heavy chains: alpha,
delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and
three (in alpha, delta and gamma) or four (in epsilon and mu) constant
domains (CH1 to CH4).

15

The major histocompatibility complex (MHC) molecules are made of two chains.
In class I [2] the alpha chain is composed of three extracellular domains, a
transmembrane region and a cytoplasmic tail. The beta chain (beta-2-
microglobulin) is composed of a single extracellular domain. In class II [3],
20 both the alpha and the beta chains are composed of two extracellular domains,
a transmembrane region and a cytoplasmic tail.

It is known [4,5] that the Ig constant chain domains and a single
extracellular domain in each type of MHC chains are related. These
25 homologous domains are approximately one hundred amino acids long and
include a conserved intradomain disulfide bond. A small pattern
around the C-terminal cysteine is involved in this disulfide bond which can be used to detect
these category of Ig related proteins.

30 Consensus pattern[FY]-x-C-x-[VA]-x-H-Sequences known to belong to this
class detected by the pattern: Ig heavy chains type Alpha C region : All,
in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig
heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy

chains type Gamma C region : All, in CH3 and also CH1 in some cases Ig
heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains
type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains
type Lambda C region : In all CL except rabbit. MHC class I alpha chains :

5 All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous
protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All,
in alpha-2 domains. MHC class II beta chains: All, in beta-2 domains.

[1]

10 Gough N.
Trends Biochem. Sci. 6:203-205(1981).

[2]

Klein J., Figueroa F.
Immunol. Today 7:41-44(1986).

15 [3]

Figueroa F., Klein J.
Immunol. Today 7:78-81(1986).

[4]

20 Orr H.T., Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.
Nature 282:266-270(1979).

[5]

Cushley W., Owen M.J.
Immunol. Today 4:88-92(1983).

[6]

25 Beck S., Barrel B.G.
Nature 331:269-272(1988).

818. (IGFBP) Insulin-like growth factor binding proteins signature
PROSITE cross-reference(s): PS00222; IGF_BINDING

30

The insulin-like growth factors (IGF-I and IGF-II) bind to specific binding
proteins in extracellular fluids with high affinity [1,2,3]. These IGF-binding
proteins (IGFBP) prolong the half-life of the IGFs and have been shown to

either inhibit or stimulate the growth promoting effects of the IGFs on cells culture. They seem to alter the interaction of IGFs with their cell surface receptors. There are at least six different IGFBPs and they are structurally related.

5

The following growth-factor inducible proteins are structurally related to IGFBPs and could function as growth-factor binding proteins [4,5]:

- Mouse protein *cyr61* and its probable chicken homolog, protein CEF-10.
- 10 - Human connective tissue growth factor (CTGF) and its mouse homolog, protein FISP-12.
- Vertebrate protein NOV.

As a signature pattern a conserved cysteine-rich region located in the N-terminal
15 section of these proteins is used.

Consensus pattern G-C-[GS]-C-C-x(2)-C-A-x(6)-C

Sequences known to belong to this class detected by the pattern ALL, except
for IGFBP-6's.

20

[1]

Rechler M.M.

Vitam. Horm. 47:1-114(1993).

[2]

25 Shimasaki S., Ling N.

Prog. Growth Factor Res. 3:243-266(1991).

[3]

Clemmons D.R.

Trends Endocrinol. Metab. 1:412-417(1990).

30

[4]

Bradham D.M., Igarashi A., Potter R.L., Grotendorst G.R.

J. Cell Biol. 114:1285-1294(1991).

[5]

Maloisel V., Martinerie C., Dambrine G., Plassiart G., Brisac M., Crochet J., Perbal B.

Mol. Cell. Biol. 12:10-21(1992).

- 5 819. LMWPc : Low molecular weight phosphotyrosine protein phosphatase
Number of members: 34

[1]Medline: 94329182, The crystal structure of a low-molecular-weight phosphotyrosine protein phosphatase. Su XD, Taddei N, Stefani M, Ramponi G, Nordlund P; Nature
10 1994;370:575-578.

820. (myosin_head) ATP/GTP-binding site motif A (P-loop)
PROSITE cross-reference(s): PS00017; ATP_GTP_A

15 From sequence comparisons and crystallographic data analysis it has been shown
[1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP
share a number of more or less conserved sequence motifs. The best conserved
of these motifs is a glycine-rich region, which typically forms a flexible
loop between a beta-strand and an alpha-helix. This loop interacts with one of
20 the phosphate groups of the nucleotide. This sequence motif is generally
referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].

There are numerous ATP- or GTP-binding proteins in which the P-loop is found.
A number of protein families for which the relevance of the
25 presence of such motif has been noted is listed below:

- ATP synthase alpha and beta subunits (see <PDOC00137>).
- Myosin heavy chains.
- Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>).
- 30 - Dynamins and dynamin-like proteins (see <PDOC00362>).
- Guanylate kinase (see <PDOC00670>).
- Thymidine kinase (see <PDOC00524>).
- Thymidylate kinase (see <PDOC01034>).

- Shikimate kinase (see <PDOC00868>).
- Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>).
- ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>).
- 5 - DNA and RNA helicases [8,9,10].
- GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.).
- Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).
- Nuclear protein ran (see <PDOC00859>).
- ADP-ribosylation factors family (see <PDOC00781>).
- 10 - Bacterial dnaA protein (see <PDOC00771>).
- Bacterial recA protein (see <PDOC00131>).
- Bacterial recF protein (see <PDOC00539>).
- Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.).
- DNA mismatch repair proteins mutS family (See <PDOC00388>).
- 15 - Bacterial type II secretion system protein E (see <PDOC00567>).

Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

25

Consensus pattern[AG]-x(4)-G-K-[ST]

[1]

Walker J.E., Saraste M., Runswick M.J., Gay N.J.

30 EMBO J. 1:945-951(1982).

[2]

Moller W., Amons R.

FEBS Lett. 186:1-7(1985).

[3]

Fry D.C., Kuby S.A., Mildvan A.S.

Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

[4]

5 Dever T.E., Glynias M.J., Merrick W.C.

Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5]

Saraste M., Sibbald P.R., Wittinghofer A.

Trends Biochem. Sci. 15:430-434(1990).

10 [6]

Koonin E.V.

J. Mol. Biol. 229:1165-1174(1993).

[7]

Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher

15 M.P.

J. Bioenerg. Biomembr. 22:571-592(1990).

[8]

Hodgman T.C.

Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

20 [9]

Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K.,

Schnier J., Slonimski P.P.

Nature 337:121-122(1989)

[10]

25 Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M.

Nucleic Acids Res. 17:4713-4730(1989).

821. PE: PE family

This family named after a PE motif near to the amino terminus of the domain. The PE family

30 of proteins all contain an amino-terminal region of about 110 amino acids. The carboxyl

terminus of this family are variable and fall into several classes. The largest class of PE

proteins is the highly repetitive PGRS class which have a high glycine content. The function

of these proteins is uncertain but it has been suggested that they may be related to antigenic variation of *Mycobacterium tuberculosis* [1]. Number of members: 88

[1] Medline: 98295987. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al; Nature 1998;393:537-544.

822. (RNB) Ribonuclease II family signature

PROSITE cross-reference(s): PS01175; RIBONUCLEASE_II

On the basis of sequence similarities, the following bacterial and eukaryotic proteins seem to form a family:

- *Escherichia coli* and related bacteria ribonuclease II (EC 3.1.13.1) (RNase II) (gene *rnb*) [1]. RNase II is an exonuclease involved in mRNA decay. It degrades mRNA by hydrolyzing single-stranded polyribonucleotides processively in the 3' to 5' direction.
- Bacterial protein *vacB*. In *Shigella flexneri*, *vacB* has been shown to be required for the expression of virulence genes at the posttranscriptional level.
- Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell cycle G1 phase.
- Yeast protein DIS3 [2], which binds to ran (GSP1) and enhances the nucleotide-releasing activity of RCC1 on ran.
- Fission yeast protein *dis3*, which is implicated in mitotic control.
- *Neurospora crassa* *cyt-4*, a mitochondrial protein required for RNA 5' and 3' end processing and splicing.
- Yeast protein MSU1, which is involved in mitochondrial biogenesis.
- *Synechocystis* strain PCC 6803 protein *zam* [3], which control resistance to the carbonic anhydrase inhibitor acetazolamide.
- *Caenorhabditis elegans* hypothetical protein F48E8.6.

The size of these proteins range from 644 residues (rnb) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [4]. It is possible that this domain plays a role in a

- 5 putative exonuclease function that would be common to all these proteins. A signature pattern was developed based on the core of this conserved domain.

Consensus pattern[HI]-[FYE]-[GSTAM][GSTAM SEQ ID NO:32]-[LIVM][LIVM SEQ ID
NO:4]-x(4,5)-Y-[STAL][STAL SEQ ID NO:471]-x-[FWVAC][FWVAC SEQ ID
10 NO:717]-[TV]-
[SA]-P-[LIVMA][LIVMA SEQ ID NO:30]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ]

[1]

Zilhao R., Camelo L., Arraiano C.M.

- 15 Mol. Microbiol. 8:43-51(1993).

[2]

Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N.,
Yanagida M., He X., Mueller U., Sazer S., Nishimoto T.
EMBO J. 15:5595-5605(1996).

- 20 [3]

Beuf L., Bedu S., Cami B., Joset F.
Plant Mol. Biol. 27:779-788(1995).

[4]

Mian I.S.

- 25 Nucleic Acids Res. 25:3187-3195(1997).

823. Src homology 2 (SH2) domain profile

PROSITE cross-reference(s): PS50001; SH2

- 30 The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as

regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].

- 5 The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].

So far, SH2 domains have been identified in the following proteins:

10

- Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases.
- Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. Two
- 15 copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>).
- Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit.
- Some vertebrate and invertebrate protein-tyrosine phosphatases.
- Mammalian Ras GTPase-activating protein (GAP).
- 20 - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, Caenorhabditis elegans sem-5 and Drosophila DRK.
- Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family.
- 25 - Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc.
- STAT proteins (signal transducers and activators of transcription).
- Chicken tensin.
- Yeast transcriptional control protein SPT6.

30

The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.

[1]

Sadowski I., Stone J.C., Pawson T.
Mol. Cell. Biol. 6:4396-4408(1986).

5

[2]

Russel R.B., Breed J., Barton G.J.
FEBS Lett. 304:15-20(1992).

[3]

10

Marangere L.E.M., Pawson T.
J. Cell Sci. Suppl. 18:97-104(1994).

[4]

Pawson T., Schlessinger J.
Curr. Biol. 3:434-442(1993).

[5]

15

Mayer B.J., Baltimore D.
Trends Cell. Biol. 3:8-13(1993).

[6]

Pawson T.
Nature 373:573-580(1995).

20

[7]

Kuriyan J., Cowburn D.
Curr. Opin. Struct. Biol. 3:828-837(1993).

824. Sulfate transporters signature

25

PROSITE cross-reference(s): PS01130; SULFATE_TRANSP

A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:

30

- Neurospora crassa sulfate permease II (gene cys-14).
- Yeast sulfate permeases (genes SUL1 and SUL2).
- Rat sulfate anion transporter 1 (SAT-1).

701

- Mammalian DTDST, a probable sulfate transporter which, in Human, is involved in the genetic disease, diastrophic dysplasia (DTD).
- Sulfate transporters 1, 2 and 3 from the legume *Stylosanthes hamata*.
- 5 - Human pendrin (gene PDS), which is involved in a number of hearing loss genetic diseases.
- Human protein DRA (Down-Regulated in Adenoma).
- Soybean early nodulin 70.
- Escherichia coli hypothetical protein ychM.
- 10 - Caenorhabditis elegans hypothetical protein F41D9.5.

As expected by their transport function, these proteins are highly hydrophobic and seem to contain about 12 transmembrane domains. The best conserved region seems to be located in the second transmembrane region and is used as a signature pattern.

Consensus pattern[PAV]-x-Y-[GS]-L-Y-~~[STAG]~~[STAG SEQ ID NO:20](2)-x(4)-~~[LIVFYA]~~[LIVFYA SEQ ID NO:718]-~~[LIVST]~~[LIVST SEQ ID NO:474]-[YI]-x(3)-[GA]-[GST]-S-[KR]

20

[1]

Sandal N.N., Marcker K.A.

Trends Biochem. Sci. 19:19-19(1994).

[2]

25 Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T.

Mol. Gen. Genet. 247:709-715(1995).

825. TYA: TYA transposon protein

Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB.

30

The TYA protein is analogous to the gag protein of retroviruses. TYA a is cleaved to form 46kd protein which can form mature virion like particles [1]. Number of members: 59

[1] Medline: 97404699. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Kingsman AJ, Fuller SD, Saibil HR; J Virol 1997;71:6863-6868.

5 826. Aldolase_II

Class II Aldolase and Adducin N-terminal domain.

-!- This family includes class II aldolases and adducins which have not been ascribed any enzymatic function. Number of members: 37

10 References:

[1] Medline: 93294819. The spatial structure of the class II L-fucose-1-phosphate aldolase from Escherichia coli. Dreyer MK, Schulz GE; J Mol Biol 1993;231:549-553.

[2] Medline: 96256522. Catalytic mechanism of the metal-dependent fucose aldolase from Escherichia coli as derived from the structure. Dreyer MK, Schulz GE; J Mol Biol

15 1996;259:458-466.

827. CBD_2

-!- Two tryptophan residues are involved in cellulose binding.

-!- Cellulose binding domain found in bacteria. Number of members: 51

20

References:

[1] Medline: 95284032. Solution structure of a cellulose-binding domain from Cellulomonas fimi by nuclear magnetic resonance spectroscopy. Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS; Biochemistry

25 1995;34:6993-7009.

828. P

A unique feature of the eukaryotic subtilisin-like proprotein convertases is the presence of an additional highly conserved sequence of approximately 150 residues (P domain) located immediately downstream of the catalytic domain.

30

Number of members: 91

References:

[1] Medline: 94252314. A C-terminal domain conserved in precursor processing proteases is required for intramolecular N-terminal maturation of pro-Kex2 protease. Gluschankof P, Fuller RS; EMBO J 1994;13:2280-2288.

[2] Medline: 98225190. Regulatory roles of the P domain of the subtilisin-like prohormone convertases. Zhou A, Martin S, Lipkind G, LaMendola J, Steiner DF; J Biol Chem 1998;273:11107-11114.

829. Uncharacterized protein family UPF0020 signature

PROSITE cross-reference(s): PS01261; UPF0020

The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Escherichia coli hypothetical protein ycbY and HI0116/15, the corresponding Haemophilus influenzae protein.

- Bacillus subtilis hypothetical protein ypsC.

- Synechocystis strain PCC 6803 hypothetical protein slr0064.

- Methanococcus jannaschii hypothetical proteins MJ0438 and MJ0710.

These are hydrophilic proteins of from 40 Kd to about 80 Kd. They can be picked up in the database by the following pattern.

Consensus pattern D-P-[LIVMF]-[LIVMF SEQ ID NO:2]-C-G-[ST]-G-x(3)-[LI]-E

References:

[1] Bairoch A. Unpublished observations (1997).

830. Uncharacterized protein family UPF0031 signatures

PROSITE cross-reference(s): PS01049; UPF0031_1; PS01050; UPF0031_2

The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Yeast chromosome XI hypothetical protein YKL151c.

- Caenorhabditis elegans hypothetical protein R107.2.

704

- Escherichia coli hypothetical protein yjeF.
- Bacillus subtilis hypothetical protein yxkO.
- Helicobacter pylori hypothetical protein HP1363.
- Mycobacterium tuberculosis hypothetical protein MtCY77.05c.
- 5 - Mycobacterium leprae hypothetical protein B229_C2_201.
- Synechocystis strain PCC 6803 hypothetical protein sll1433.
- Methanococcus jannaschii hypothetical protein MJ1586.

These are proteins of about 30 to 40 Kd whose central region is well

10 conserved. They can be picked up in the database by the following patterns.

Consensus pattern[SAV]-[IVW]-[LVA]-[LIV]-G-[PNS]-G-L-[GP]-x-[DENQT][DENQT
SEQ ID NO:719]

Consensus pattern[GA]-G-x-G-D-[TV]-[LT]-[STA]-G-x-[LIVM][LIVM SEQ ID NO:4]

15

831. (ACOX)

Acyl-CoA oxidase

This is a family of Acyl-CoA oxidases EC:1.3.3.6. Acyl-coA oxidase converts acyl-CoA into
20 trans-2-enoyl-CoA [1].

Number of members: 39

[1] Hayashi H, De Bellis L, Yamaguchi K, Kato A, Hayashi M, Nishimura M; Medline:

25 98192624. "Molecular characterization of a glyoxysomal long chain acyl-CoA oxidase that is
synthesized as a precursor of higher molecular mass in pumpkin." J Biol Chem
1998;273:8301-8307.

30

832. (AICARFT_IMPCHas)

AICARFT/IMPCHase bienzyme

705

This is a family of bifunctional enzymes catalysing the last steps in de novo purine biosynthesis. The bifunctional enzyme is found in both prokaryotes and eukaryotes. The second last step is catalysed by 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase EC:2.1.2.3 (AICARFT), this enzyme catalyses the formylation of AICAR with 10-formyl-tetrahydrofolate to yield FAICAR and tetrahydrofolate [1]. The last step is catalysed by IMP (Inosine monophosphate) cyclohydrolase EC:3.5.4.10 (IMPCHase), cyclizing FAICAR (5-formylaminoimidazole-4-carboxamide ribonucleotide) to IMP [1].

Number of members: 22

[1] Akira T, Komatsu M, Nango R, Tomooka A, Konaka K, Yamauchi M, Kitamura Y, Nomura S, Tsukamoto I; Medline: 97473523 "Molecular cloning and expression of a rat cDNA encoding 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase" [published erratum appears in Gene 1998 Feb 27;208(2):337] Gene 1997;197:289-293.

[2] Rayl EA, Moroson BA, Beardsley GP; Medline: 96147205 "The human purH gene product, 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase. Cloning, sequencing, expression, purification, kinetic analysis, and domain mapping." J Biol Chem 1996;271:2225-2233.

833. (AOX)

Alternative oxidase

The alternative oxidase is used as a second terminal oxidase in the mitochondria, electrons are transferred directly from reduced ubiquinol to oxygen forming water [2]. This is not coupled to ATP synthesis and is not inhibited by cyanide, this pathway is a single step process [1]. In rice the transcript levels of the alternative oxidase are increased by low temperature [1].

Number of members: 27

[1] Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A; Medline: 98086211 "Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature." Gene 1997;203:121-129.

5 [2] Li Q, Ritzel RG, McLean LL, McIntosh L, Ko T, Bertrand H, Nargang FE; Medline: 96366413 "Cloning and analysis of the alternative oxidase gene of *Neurospora crassa*." Genetics 1996;142:129-140.

10 834. (APH)

Protein kinases signatures and profile

Cross-reference(s): PS00107; PROTEIN_KINASE_ATP, PS00108;
PROTEIN_KINASE_ST, PS00109; PROTEIN_KINASE_TYR, PS50011;
15 PROTEIN_KINASE_DOM

Eukaryotic protein kinases [1 to 5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein
20 kinases. Two of these regions have been selected to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the
25 enzyme [6]; two signature patterns were derived for that region: one specific for serine/threonine kinases and the other for tyrosine kinases. A profile was developed which is based on the alignment in [1] and covers the entire catalytic domain.

Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH][FYWMGSTNH SEQ ID
30 NO:441)]-[SGA]-{PW}-[LIVCAT][LIVCAT SEQ ID NO:442)]-{PD}-x-
[GSTACLIVMFY][GSTACLIVMFY SEQ ID NO:443)]-x(5,18)-
[LIVMFYWCSTAR][LIVMFYWCSTAR SEQ ID NO:444)]-[AIVP][AIVP SEQ ID
NO:445)]-[LIVMFAGCKR][LIVMFAGCKR SEQ ID NO:446)]-K [K binds ATP]

Sequences known to belong to this class detected by the pattern the majority of known protein kinases but it fails to find a number of them, especially viral kinases which are quite divergent in this region and are completely missed by this pattern.

5

Consensus pattern: [LIVMFYC][LIVMFYC SEQ ID NO:6]]-x-[HY]-x-D-
[LIVMFY][LIVMFY SEQ ID NO:18]]-K-x(2)-N-[LIVMFYCT][LIVMFYCT SEQ ID
NO:447]](3) [D is an active site residue]

10 Sequences known to belong to this class detected by the pattern. Most serine/ threonine specific protein kinases with 10 exceptions (half of them viral kinases) and also Epstein-Barr virus BGLF4 and Drosophila ninaC which have respectively Ser and Arg instead of the conserved Lys and which are therefore detected by the tyrosine kinase specific pattern described below.

15

Consensus pattern: [LIVMFYC][LIVMFYC SEQ ID NO:6]]-x-[HY]-x-D-
[LIVMFY][LIVMFY SEQ ID NO:18]]-[RSTAC][RSTAC SEQ ID NO:448]]-x(2)-N-
[LIVMFYC][LIVMFYC SEQ ID NO:6]](3) [D is an active site residue] tyrosine specific
 protein kinases with the exception of human ERBB3 and mouse blk. This pattern will also
 20 detect most bacterial aminoglycoside phosphotransferases [8,9] and herpesviruses ganciclovir
 kinases [10]; which are proteins structurally and evolutionary related to protein kinases.
 Sequences known to belong to this class detected by the profile ALL, except for three viral
 kinases. This profile also detects receptor guanylate cyclases (see <PDOC00430>) and 2-5A-
 dependent ribonucleases. Sequence similarities between these two families and the eukaryotic
 25 protein kinase family have been noticed before. It also detects Arabidopsis thaliana kinase-
 like protein TMKL1 which seems to have lost its catalytic activity.

Note if a protein analyzed includes the two protein kinase signatures, the probability of it
 being a protein kinase is close to 100%. Note eukaryotic-type protein kinases have also been
 30 found in prokaryotes such as Myxococcus xanthus [11] and Yersinia pseudotuberculosis.
 Note the patterns shown above has been updated since their publication in [7]. Note this
 documentation entry is linked to both signature patterns and a profile. As the profile is much

more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

References

- 5 [1] Hanks S.K., Hunter T., FASEB J. 9:576-596(1995).
- [2] Hunter T., Meth. Enzymol. 200:3-37(1991).
- [3] Hanks S.K., Quinn A.M., Meth. Enzymol. 200:38-62(1991).
- [4] Hanks S.K., Curr. Opin. Struct. Biol. 1:369-383(1991).
- [5] Hanks S.K., Quinn A.M., Hunter T., Science 241:42-52(1988).
- 10 [6] Knighton D.R., Zheng J., Ten Eyck L.F., Ashford V.A., Xuong N.-H., Taylor, S.S., Sowadski J.M., Science 253:407-414(1991).
- [7] Bairoch A., Claverie J.-M., Nature 331:22(1988).
- [8] Benner S., Nature 329:21-21(1987).
- [9] Kirby R., J. Mol. Evol. 30:489-492(1992).
- 15 [10] Littler E., Stuart A.D., Chee M.S., Nature 358:160-162(1992).
- [11] Munoz-Dorado J., Inouye S., Inouye M., Cell 67:995-1006(1991).

835. (Asp_Glu_race)

20 Aspartate and glutamate racemases signatures

Cross-reference(s) PS00923; ASP_GLU_RACEMASE_1 PS00924;
ASP_GLU_RACEMASE_2

- 25 Aspartate racemase (EC 5.1.1.13) and glutamate racemase (EC 5.1.1.3) are two evolutionary related bacterial enzymes that do not seem to require a cofactor for their activity [1].
Glutamate racemase, which interconverts L-glutamate into D-glutamate, is required for the biosynthesis of peptidoglycan and some peptide-based antibiotics such as gramicidin S. In addition to characterized aspartate and glutamate racemases, this family also includes a
- 30 hypothetical protein from Erwinia carotovora and one from Escherichia coli (ygeA). Two conserved cysteines are present in the sequence of these enzymes. They are expected to play a role in catalytic activity by acting as bases in proton abstraction from the substrate.
Signature patterns were developed for both cysteines.

Consensus pattern: [IVA]-[LIVM][LIVM SEQ ID NO:4]-x-C-x(0,1)-N-[ST]-[MSA]-[STH]-
[LIVFYSTANK][LIVFYSTANK SEQ ID NO:720)]

5 Consensus pattern: [LIVM][LIVM SEQ ID NO:4]](2)-x-[AG]-C-T-[DEH]-
[LIVMFY][LIVMFY SEQ ID NO:18)]-[PNGRS][PNGRS SEQ ID NO:721)]-x-
[LIVM][LIVM SEQ ID NO:4)]

[1] Gallo K.A., Knowles J.R., Biochemistry 32:3981-3990(1993).

10

836. (ATP-sulfurylase)

ATP-sulfurylase

15 This family consists of ATP-sulfurylase or sulfate adenylyltransferase EC:2.7.7.4 some of
which are part of a bifunctional polypeptide chain associated with adenosyl phosphosulphate
(APS) kinase APS_kinase. Both enzymes are required for PAPS (phosphoadenosine-
phosphosulfate) synthesis from inorganic sulphate [2]. ATP sulfurylase catalyses the
synthesis of adenosine-phosphosulfate APS from ATP and inorganic sulphate [1].

20

Number of members: 37

[1] Kurima K, Warman ML, Krishnan S, Domowicz M, Krueger RC Jr, Deyrup A, Schwartz
NB; Medline: 98337975 "A member of a family of sulfate-activating enzymes causes murine
25 brachymorphism" [published erratum appears in Proc Natl Acad Sci U S A 1998 Sep
29;95(20):12071] Proc Natl Acad Sci U S A 1998;95:8681-8685.

[2] Rosenthal E, Leustek T; Medline: 96096529 "A multifunctional Urechis caupo protein,
PAPS synthetase, has both ATP sulfurylase and APS kinase activities." Gene 1995;165:243-
30 248.

837. (ATP-synt_F)

ATP synthase (F/14-kDa) subunit

This family includes 14-kDa subunit from vATPases [1], which is in the peripheral catalytic part of the complex [2]. The family also includes archaebacterial ATP synthase subunit F [3].

5

Number of members: 23

[1] Guo Y, Kaiser K, Wiczorek H, Dow JA; Medline: 96269411 "The *Drosophila melanogaster* gene *vha14* encoding a 14-kDa F-subunit of the vacuolar ATPase." *Gene* 1996;172:239-243.

10

[2] Peng SB, Crider BP, Tsai SJ, Xie XS, Stone DK; Medline: 96216416 "Identification of a 14-kDa subunit associated with the catalytic sector of clathrin-coated vesicle H⁺-ATPase." *J Biol Chem* 1996;271:3324-3327.

[3] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." *J Biol Chem* 1996;271:18843-18852.

15

838. (CBD_4)

20

Starch binding domain

Number of members: 48

839. (CbiX)

25

The function of CbiX is uncertain, however it is found in cobalamin biosynthesis operons and so may have a related function. Some CbiX proteins contain a striking histidine-rich region at their C-terminus, which suggests that it might be involved in metal chelation [1].

30

Number of members: 6

[1] Raux E, Lanois A, Warren MJ, Rambach A, Thermes C; Medline: 98416126 "Cobalamin (vitamin B12) biosynthesis: identification and characterization of a *Bacillus megaterium* cobI operon." *Biochem J* 1998;335:159-166.

5

840. (Complex1_51K)

Respiratory-chain NADH dehydrogenase 51 Kd subunit signatures Cross-reference(s)
PS00644; COMPLEX1_51K_1 PS00645; COMPLEX1_51K_2

10

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this
15 bioenergetic enzyme complex there is one with a molecular weight of 51 Kd (in mammals), which is the second largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind to NAD, FMN, and a 2Fe-2S cluster.

The 51 Kd subunit is highly similar to [3,4]:

20

- Subunit alpha of *Alcaligenes eutrophus* NAD-reducing hydrogenase (gene *hoxF*) which also binds to NAD, FMN, and a 2Fe-2S cluster.
- Subunit NQO1 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- Subunit F of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoF*).

25

The 51 Kd subunit and the bacterial hydrogenase alpha subunit contains three regions of sequence similarities. The first one most probably corresponds to the NAD-binding site, the second to the FMN-binding site, and the third one, which contains three cysteines, to the iron-sulfur binding region. Signature patterns have been developed for the FMN-binding and for the 2Fe-2S binding regions.

30

Consensus pattern: G-[AM]-G-[AR]-Y-[LIVM][LIVM SEQ ID NO:4]-C-G-[DE](2)-[STA](2)-[LIM](2)-[EN]- S

Consensus pattern: E-S-C-G-x-C-x-P-C-R-x-G [The three C's are putative 2Fe-2S ligands]

[1] Ragan C.I., Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D., Eur. J. Biochem. 197:563-576(1991).

[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

5 [4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H., J. Mol. Biol. 233:109-122(1993).

841. (DAP_epimerase)

10 Diaminopimelate epimerase signature

Cross-reference(s) PS01326; DAP_EPIMERASE

15 Diaminopimelate epimerase (EC 5.1.1.7) catalyzes the isomeriazation of L,L- to D,L-meso-diaminopimelate in the biosynthetic pathway leading from aspartate to lysine. This enzyme is a protein of about 30 Kd. Two conserved cysteines seem [1] to function as the acid and base in the catalytic mechanism. As a signature pattern, the region surrounding the first of these two active site cysteines were selected.

20 Consensus pattern: N-x-D-G-S-x(4)-C-G-N-[GA]-x-R [C is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for an Anabaena dapF which has a Ser instead of the active site Cys.

[1] Cirilli M., Zheng R., Scapin G., Blanchard J.S., Biochemistry 37:16452-16458(1998).

25

842. (DNA_gyraseB_C)

DNA topoisomerase II signature

Cross-reference(s) PS00177; TOPOISOMERASE_II

30 DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in

713

African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as

5 topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following

10 representation:

<-----About-1400-residues----->

[-----Protein 39-*-----]	[----Protein 52----]	Phage T4
[-----gyrB-----*-----]	[-----gyrA-----]	Prokaryote II
	Archaeobacteria	
[-----parE-----*-----]	[-----parD-----]	Prokaryote IV
[-----*-----]		Eukaryote and
	ASF	

20 '*': Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

25 Consensus pattern: [LIVMA][LIVMA SEQ ID NO:30]-x-E-G-[DN]-S-A-x-[STAG][STAG SEQ ID NO:20]

[1] Sternglanz R., Curr. Opin. Cell Biol. 1:533-535(1990).

30 [2] Bjornsti M.-A., Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A., Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J., Trends Biochem. Sci. 20:156-160(1995).

843. (DUF16)

Protein of unknown function

- 5 The function of this protein is unknown. It appears to only occur in *Mycoplasma pneumoniae*.

Number of members: 26

- 10 [1] Himmelreich R, Hilbert H, Plagens H, Pirkel E, Li BC, Herrmann R; Medline: 97105885
“Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.”
Nucleic Acids Res 1996;24:4420-4449.

15 844. (DUF21)

Domain of unknown function

- 20 This transmembrane region has no known function. Many of the sequences in this family are
annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not
contain this domain. This domain is found in the N-terminus of the proteins adjacent to two
intracellular CBS domains CBS.

Number of members: 42

25

845. (DUF56)

Integral membrane protein

30

The members of this family are putative integral membrane proteins. The function of the family is unknown, however the family includes Sec59 from yeast. Sec59 is a dolichol

715

kinase EC:2.7.1.108, but it is not clear if the enzymatic activity resides in this region or its N terminal region.

Number of members: 13

5

846. (DUF94)

Domain of unknown function

10

The function of this domain is unknown. It is found in both eukaryotes and archaeobacteria. The alignment contains a completely conserved aspartate residue that may be functionally important. The eukaryotic domains contains three conserved cysteines and a histidine that might be metal binding, however these are absent in the archaeobacterial proteins.

15

Number of members: 9

847. (FF)

20

FF domain

This domain may be involved in protein-protein interaction [1].

25

Number of members: 42

[1] Bedford MT, Leder P; Medline: 99322199 "The FF domain: a novel motif that often accompanies WW domains." Trends Biochem Sci 1999;24:264-265.

30

848. (FLO_LFY)

Floricaula / Leafy protein

This family consists of various plant development proteins which are homologues of floricaula (FLO) and Leafy (LFY) proteins which are floral meristem identity proteins. Mutations in the sequences of these proteins affect flower and leaf development.

5 Number of members: 16

[1] Hofer J, Turner L, Hellens R, Ambrose M, Matthews P, Michael A, Ellis N; Medline: 97411151 "UNIFOLIATA regulates leaf and flower morphogenesis in pea." Curr Biol 1997;7:581-587.

10 [2] Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM; Medline: 92274452 "LEAFY controls floral meristem identity in Arabidopsis." Cell 1992;69:843-859.

849. (G-patch)

15 G-patch domain

This domain is found in a number of RNA binding proteins, and is also found in proteins that contain RNA binding domains. This suggests that this domain may have an RNA binding function. This domain has seven highly conserved glycines.

20

Number of members: 47

[1] Aravind L, Koonin EV; Medline: 10470032 "G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins." Trends Biochem Sci 1999;24:342-344.

25

850. (Gram-ve_porins)

General diffusion Gram-negative porins signature

30

Cross-reference(s) PS00576; GRAM_NEG_PORIN

The outer membrane of Gram-negative bacteria acts as a molecular filter for hydrophilic compounds. Proteins, known as porins [1], are responsible for the 'molecular sieve' properties

717

of the outer membrane. Porins form large water- filled channels which allows the diffusion of hydrophilic molecules into the periplasmic space. Some porins form general diffusion channels that allows any solutes up to a certain size (that size is known as the exclusion limit) to cross the membrane, while other porins are specific for a solute and contain a binding site for that solute inside the pores (these are known as selective porins). As porins are the major outer membrane proteins, they also serve as receptor sites for the binding of phages and bacteriocins. General diffusion porins generally assemble as trimer in the membrane and the transmembrane core of these proteins is composed exclusively of beta strands [2]. It has been shown [3] that a number of general porins are evolutionary related, these porins are:

- Enterobacteria phoE.
- Enterobacteria ompC.
- Enterobacteria ompF.
- Enterobacteria nmpC.
- Bacteriophage PA-2 LC.
- Neisseria PI.A.
- Neisseria PI.B.

As a signature pattern a conserved region was selected, located in the C-terminal part of these proteins, which spans two putative transmembrane beta strands.

Consensus pattern: ~~[LIVMFY]~~~~[LIVMFY SEQ ID NO:18]~~-x(2)-G-x(2)-Y-x-F-x-K-x(2)-
[SN]-~~[STAV]~~~~[STAV SEQ ID NO:105]~~-~~[LIVMFYW]~~~~[LIVMFYW SEQ ID NO:26]~~- V

[1] Benz R., Bauer K., Eur. J. Biochem. 176:1-19(1988).

[2] Jap B.K., Walian P.J., Q. Rev. Biophys. 23:367-403(1990).

[3] Jeanteur D., Lakey J.H., Pattus F., Mol. Microbiol. 5:2153-2164(1991).

851. (HlyD)

HlyD family secretion proteins signature

Cross-reference(s) PS00543; HLYD_FAMILY

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, require the help of two or more proteins for their secretion across the cell envelope. Amongst which a protein belonging to the ABC transporters family (see the relevant entry <PDOC00185>) and a protein belonging to a family which is currently composed [1 to 5] of the following members:

Gene	Species	Protein which is exported
------	---------	---------------------------

Gene	Species	Protein which is exported
------	---------	---------------------------

hlyD	Escherichia coli	Hemolysin
------	------------------	-----------

appD	A.pleuropneumoniae	Hemolysin
------	--------------------	-----------

lcnD	Lactococcus lactis	Lactococcin A
------	--------------------	---------------

lktD	A.actinomycetemcomitans	Leukotoxin
------	-------------------------	------------

	Pasteurella haemolytica	
--	-------------------------	--

rtxD	A.pleuropneumoniae	Toxin-III
------	--------------------	-----------

cyaD	Bordetella pertussis	Calmodulin-sensitive adenylate cyclase-hemolysin (cyclolysin)
------	----------------------	---

cvaA	Escherichia coli	Colicin V
------	------------------	-----------

prtE	Erwinia chrysanthemi	Extracellular proteases B and C
------	----------------------	---------------------------------

aprE	Pseudomonas aeruginosa	Alkaline protease
------	------------------------	-------------------

emrA	Escherichia coli	Drugs and toxins
------	------------------	------------------

yjcR	Escherichia coli	Unknown
------	------------------	---------

These proteins are evolutionary related and consist of from 390 to 480 amino acid residues.

They seem to be anchored in the inner membrane by a N-terminal transmembrane region.

Their exact role in the secretion process is not yet known. The C-terminal section of these proteins is the best conserved region; a signature pattern from that region was derived.

Consensus pattern: [LIVM][LIVM SEQ ID NO:4]]-x(2)-G-[LM]-x(3)-[STGAV][STGAV SEQ ID NO:722]]-x-[LIVMTF][LIVMT SEQ ID NO:1]]-x-[LIVMTF][LIVMT SEQ ID NO:1]]-[GE]-x-[KR]-x-[LIVMFYW][LIVMFYW SEQ ID NO:26]](2)-x-[LIVMFYW][LIVMFYW SEQ ID NO:26]](3)

Sequences known to belong to this class detected by the pattern ALL, except for emrA and yjcR.

References:

- [1] Gilson L., Mahanty H.K., Kolter R., EMBO J. 9:3875-3884(1990).
- [2] Letoffe S., Delepelaire P., Wandersman C., EMBO J. 9:1375-1382(1990).
- [3] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L., Appl. Environ.
- 5 Microbiol. 58:1952-1961(1992).
- [4] Duong F., Lazdunski A., Cami B., Murgier M., Gene 121:47-54(1992).
- [5] Lewis K., Trends Biochem. Sci. 19:119-123(1994).

10 852. (IBR)

In Between Ring fingers

The IBR (In Between Ring fingers) domain is found to occur between pairs of ring fingers (zf-C3HC4). The function of this domain is unknown. This domain has also been called the

15 C6HC domain and DRIL (for double RING finger linked) domain [2].

Number of members: 25

[1] Morett E, Bork P; Medline: 10366851 "A novel transactivation domain in parkin."Trends Biochem Sci 1999;24:229-231.

20 [2] van der Reijden BA, Erpelinck-Verschueren CA, Lowenberg B, Jansen JH; Medline: 99349709 "TRIADs: a new class of proteins with a novel cysteine-rich signature." Protein Sci 1999;8:1557-1561.

25 853. (IPPT)

IPP transferase

[1] Durand JM, Bjork GR, Kuwae A, Yoshikawa M, Sasakawa C; Medline: 97440126 "The modified nucleoside 2-methylthio-N6-isopentenyladenosine in tRNA of Shigella flexneri is

30 required for expression of virulence genes." J Bacteriol 1997;179:5777-5782.

[2] Boguta M, Hunter LA, Shen WC, Gillman EC, Martin NC, Hopper AK; Medline: 94187700 "Subcellular locations of MOD5 proteins: mapping of sequences sufficient for

720

targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol." Mol Cell Biol 1994;14:2298-2306.

[3] Gillman EC, Slusher LB, Martin NC, Hopper AK; Medline: 91203856 "MOD5 translation initiation sites determine N6-isopentenyladenosine modification of mitochondrial and cytoplasmic tRNA." Mol Cell Biol 1991;11:2382-2390.

854. (KE2)

KE2 family protein

The function of members of this family is unknown, although they have been suggested to contain a DNA binding leucine zipper motif [2].

Number of members: 9

[1] Ha H, Abe K, Artzt K; Medline: 92084131 "Primary structure of the embryo-expressed gene KE2 from the mouse H-2K region." Gene 1991;107:345-346.

[2] Shang HS, Wong SM, Tan HM, Wu M; Medline: 95129859 "YKE2, a yeast nuclear gene encoding a protein showing homology to mouse KE2 and containing a putative leucine-zipper motif." Gene 1994;151:197-201.

855. (Lipoprotein_6)

Prokaryotic membrane lipoprotein lipid attachment site

Cross-reference(s) PS00013; PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).

- *Escherichia coli* lipoprotein-34 (gene nlpB).
- *Escherichia coli* lipoprotein nlpC.
- *Escherichia coli* lipoprotein nlpD.
- *Escherichia coli* osmotically inducible lipoprotein B (gene osmB).
- 5 - *Escherichia coli* osmotically inducible lipoprotein E (gene osmE).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene pal).
- *Escherichia coli* rare lipoproteins A and B (genes rplA and rplB).
- *Escherichia coli* copper homeostasis protein cutF (or nlpE).
- *Escherichia coli* plasmids traT proteins.
- 10 - *Escherichia coli* Col plasmids lysis proteins.
- A number of *Bacillus* beta-lactamases.
- *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA).
- *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB).
- *Borrelia hermsii* variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- 15 - *Chlamydia trachomatis* outer membrane protein 3 (gene omp3).
- *Fibrobacter succinogenes* endoglucanase cel-3.
- *Haemophilus influenzae* proteins Pal and Pcp.
- *Klebsiella pullulunase* (gene pulA).
- *Klebsiella pullulunase* secretion protein pulS.
- 20 - *Mycoplasma hyorhinis* protein p37.
- *Mycoplasma hyorhinis* variant surface antigens A, B, and C (genes vlpABC).
- *Neisseria* outer membrane protein H.8.
- *Pseudomonas aeruginosa* lipopeptide (gene lppL).
- *Pseudomonas solanacearum* endoglucanase egl.
- 25 - *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
- *Rickettsia* 17 Kd antigen.
- *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
- *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
- *Treponema pallidum* 34 Kd antigen.
- 30 - *Treponema pallidum* membrane protein A (gene tmpA).
- *Vibrio harveyi* chitinase (gene chb).
- *Yersinia* virulence plasmid protein yscJ.

- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification were derived.

Consensus pattern: {DERK}{DERK SEQ ID NO:354}}(6)-
 [LIVMEFWSTAG][LIVMEFWSTAG SEQ ID NO:352}}(2)-
 [LIVMEYSTAGCQ][LIVMEYSTAGCQ SEQ ID NO:353}}-[AGS]-C [C is the lipid
 attachment site] Additional rules: 1)

The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

References

- [1] Hayashi S., Wu H.C., J. Bioenerg. Biomembr. 22:451-471(1990).
- [2] Klein P., Somorjai R.L., Lau P.C.K., Protein Eng. 2:15-20(1988).
- [3] von Heijne G., Protein Eng. 2:531-534(1989).
- [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

856. (Lipoprotein_7)
 Adhesin lipoprotein

This family consists of the p50 and variable adherence-associated antigen (Vaa) adhesins from *Mycoplasma hominis*. *M. hominis* is a mycoplasma associated with human urogenital diseases, pneumonia, and septic arthritis [1]. An adhesin is a cell surface molecule that mediates adhesion to other cells or to the surrounding surface or substrate. The Vaa antigen is a 50-kDa surface lipoprotein that has four tandem repetitive DNA sequences encoding a

periodic peptide structure, and is highly immunogenic in the human host [1]. p50 is also a 50-kDa lipoprotein, having three repeats A,B and C, that may be a tetramer of 191-kDa in its native environment [2].

5 Number of members: 18

[1] Zhang Q, Wise KS; Medline: 96294788 “Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent vaa genes. “ *Infect Immun* 1996;64:2737-2744.

10 [2] Henrich B, Kitzerow A, Feldmann RC, Schaal H, Hadding U; Medline: 97047675 “Repetitive elements of the *Mycoplasma hominis* adhesin p50 can be differentiated by monoclonal antibodies.” *Infect Immun* 1996;64:4027-4034.

15 857. (MaoC_like)
MaoC like domain

The MaoC protein is found to share similarity with a wide variety of enzymes; estradiol 17 beta-dehydrogenase 4, peroxisomal hydratase-dehydrogenase-epimerase, fatty acid synthase
20 beta subunit. All these enzymes contain other domains. This domain is also present in the NodN nodulation protein N. No specific function has been assigned to this region of any of these proteins. The maoC gene is part of a operon with maoA which is involved in the synthesis of monoamine oxidase [1].

25 Number of members: 46

[1] Sugino H, Sasaki M, Azakami H, Yamashita M, Murooka Y Medline: 96235221 “A monoamine-regulated *Klebsiella aerogenes* operon containing the monoamine oxidase structural gene (maoA) and the maoC gene.” *J Bacteriol* 1992;174:2485-2492.

30

858. (MSP)
Manganese-stabilizing protein / photosystem II polypeptide

This family consists of the 33 KDa photosystem II polypeptide from the oxygen evolving complex (OEC) of plants and cyanobacteria. The protein is also known as the manganese-stabilizing protein as it is associated with the manganese complex of the OEC and may provide the ligands for the complex [1].

Number of members: 17

[1] Philbrick JB, Zilinskas BA; Medline: 88334494 "Cloning, nucleotide sequence and mutational analysis of the gene encoding the Photosystem II manganese-stabilizing polypeptide of *Synechocystis* 6803." *Mol Gen Genet* 1988;212:418-425.

859. (NAC)

[1] Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV; Medline: 99342100 "Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell." *Genome Res* 1999;9:608-628.

Number of members: 27

860. (Nop)

Putative snoRNA binding domain

This family consists of various Pre RNA processing ribonucleoproteins. The function of the aligned region is unknown however it may be a common RNA or snoRNA or Nop1p binding domain. Nop5p (Nop58p) Swiss:Q12499 from yeast is the protein component of a ribonucleoprotein protein required for pre-18s rRNA processing and is suggested to function with Nop1p in a snoRNA complex [1]. Nop56p Swiss:O00567 and Nop5p interact with Nop1p and are required for ribosome biogenesis [2]. Prp31p Swiss:p49704 is required for pre-mRNA splicing in *S. cerevisiae* [3].

Number of members: 23

[1] Wu P, Brockenbrough JS, Metcalfe AC, Chen S, Aris JP; Medline: 98298165 "Nop5p is a small nucleolar ribonucleoprotein component required for pre- 18 S rRNA processing in yeast." J Biol Chem 1998;273:16453-16463.

[2] Gautier T, Berges T, Tollervy D, Hurt E; Medline: 8038777 "Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis." Mol Cell Biol 1997;17:7088-7098.

[3] Weidenhammer EM, Singh M, Ruiz-Noriega M, Woolford JL Jr; Medline: 96184869 "The PRP31 gene encodes a novel protein required for pre-mRNA splicing in *Saccharomyces cerevisiae*." Nucleic Acids Res 1996;24:1164-1170.

861. (Nramp)

Natural resistance-associated macrophage protein

The natural resistance-associated macrophage protein (NRAMP) family consists of Nramp1, Nramp2, and yeast proteins Smf1 and Smf2. The NRAMP family is a novel family of functional related proteins defined by a conserved hydrophobic core of ten transmembrane domains [5]. This family of membrane proteins are divalent cation transporters. Nramp1 is an integral membrane protein expressed exclusively in cells of the immune system and is recruited to the membrane of a phagosome upon phagocytosis [1]. By controlling divalent cation concentrations Nramp1 may regulate the interphagosomal replication of bacteria [1]. Mutations in Nramp1 may genetically predispose an individual to susceptibility to diseases including leprosy and tuberculosis conversely this might however provide protection from rheumatoid arthritis [1]. Nramp2 is a multiple divalent cation transporter for Fe²⁺, Mn²⁺ and Zn²⁺ amongst others it is expressed at high levels in the intestine; and is major transferrin-independent iron uptake system in mammals [1]. The yeast proteins Smf1 and Smf2 may also transport divalent cations [3].

Number of members: 36

[1] Govoni G, Gros P; Medline: 98383996 "Macrophage NRAMP1 and its role in resistance to microbial infections." *Inflamm Res* 1998;47:277-284.

[2] Agranoff DD, Krishna S Medline: 98294035 "Metal ion homeostasis and intracellular parasitism." *Mol Microbiol* 1998;28:403-412.

5 [3] Pinner E, Gruenheid S, Raymond M, Gros P; Medline: 98030569 "Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a member of the mammalian natural resistance- associated macrophage protein family." *J Biol Chem* 1997;272:28933-28938.

10 [4] Cellier M, Belouchi A, Gros P; Medline: 96402487 "Resistance to intracellular infections: comparative genomic analysis of Nramp." *Trends Genet* 1996;12:201-204.

[5] Cellier M, Prive G, Belouchi A, Kwan T, Rodrigues V, Chia W, Gros P; Medline: 96036029 "Nramp defines a family of membrane proteins." *Proc Natl Acad Sci U S A* 1995;92:10089-10093.

15

862. (NTP_transf_2)

Nucleotidyltransferase domain

Members of this family belong to a large family of nucleotidyltransferases [1].

20

Number of members: 83

[1] Holm L, Sander C; Medline: 96005605 "DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily." *Trends Biochem Sci* 1995;20:345-347.

25

863. (Paramyxo_P)

Paramyxovirus P phosphoprotein

30

This family consists of paramyxovirus P phosphoprotein from sendai virus and human and bovine parainfluenza viruses. The P protein is an essential part of the viral RNA polymerase complex formed from the P and L proteins [1]. The exact role of the P protein in this complex is unknown but it is involved in multiple protein-protein interactions and binding the

polymerase complex to the nucleocapsid or ribonucleoprotein template [1]. It also appears to be important for the proper folding of the L protein [1]. The paramyxoviruses have a negative sense ssRNA genome [1].

5 Number of members: 15

[1] Bowman MC, Smallwood S, Moyer SA; Medline: 99329169 "Dissection of Individual Functions of the Sendai Virus Phosphoprotein in Transcription." J Virol 1999;73:6474-6483.

10 [2] Matsuoka Y, Curran J, Pelet T, Kolakofsky D, Ray R, Compans RW; Medline: 91237868
"The P gene of human parainfluenza virus type 1 encodes P and C proteins but not a cysteine-rich V protein." J Virol 1991;65:3406-3410.

864. (Patatin)

15

This family consists of various patatin glycoproteins from plants. The patatin protein accounts for up to 40% of the total soluble protein in potato tubers [2]. Patatin is a storage protein but it also has the enzymatic activity of lipid acyl hydrolase, catalysing the cleavage of fatty acids from membrane lipids [2].

20

Number of members: 21

[1] Banfalvi Z, Kostyal Z, Barta E; Medline: 95107249 "Solanum brevidens possesses a non-sucrose-inducible patatin gene." Mol Gen Genet 1994;245:517-522.

25 [2] Mignery GA, Pikaard CS, Park WD; Medline: 88226014 "Molecular characterization of the patatin multigene family of potato." Gene 1988;62:27-44.

865. (Pentapeptide_2)

30 Pentapeptide repeats (8 copies)

These repeats are found in many mycobacterial proteins. These repeats are most common in the PPE family of proteins, where they are found in the MPTR subfamily of PPE proteins.

The function of these repeats is unknown. The repeat can be approximately described as XNXGX, where X can be any amino acid. These repeats are similar to Pentapeptide [1], however it is not clear if these two families are structurally related.

5 Number of members: 362

[1] Bateman A, Murzin A, Teichmann SA; Medline: 98318059 "Structure and distribution of pentapeptide repeats in bacteria." Protein Sci 1998;7:1477-1480.

10 [2] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG; Medline: 98295987 "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 1998;393:537-544.

15

866. (Peptidase_C13)

Peptidase C13 family

20 This family of peptidases is known as the hemoglobinase family because it contains a globin degrading enzyme from blood parasites Swiss:P42665. However relatives are found in plants and other organisms that have other functions. Members of this family are asparaginyl peptidases [1].

25 Number of members: 26

[1] Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ; Medline: 97218252 "Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase." J Biol Chem 1997;272:8090-8098.

30

867. (Pro_dh)

Proline dehydrogenase

Number of members: 25

[1] Ling M, Allen SW, Wood JM; Medline: 95055736 "Sequence analysis identifies the proline dehydrogenase and delta 1- pyrroline-5-carboxylate dehydrogenase domains of the multifunctional *Escherichia coli* PutA protein." J Mol Biol 1994;243:950-956.

868. (PsbP)

This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria. The 23 KDa PsbP protein is required for PSII to be fully operational in vivo, it increases the affinity of the water oxidation site for Cl- and provides the conditions required for high affinity binding of Ca²⁺ [2].

Number of members: 25

[1] Rova EM, Mc Ewen B, Fredriksson PO, Styring S; Medline: 97067138 "Photoactivation and photoinhibition are competing in a mutant of *Chlamydomonas reinhardtii* lacking the 23-kDa extrinsic subunit of photosystem II." J Biol Chem 1996;271:28918-28924.

[2] Kochhar A, Khurana JP, Tyagi AK; Medline: 97191538 "Nucleotide sequence of the psbP gene encoding precursor of 23-kDa polypeptide of oxygen-evolving complex in *Arabidopsis thaliana* and its expression in the wild-type and a constitutively photomorphogenic mutant." DNA Res 1996;3:277-285.

869. (PUA)

The PUA domain named after PseudoUridine synthase and Archaeosine transglycosylase, was detected in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, a family of predicted ATPases that may be involved in RNA modification, a family of predicted archaeal and bacterial rRNA methylases. Additionally, the PUA domain

was detected in a family of eukaryotic proteins that also contain a domain homologous to the translation initiation factor eIF1/SUI1; these proteins may comprise a novel type of translation factors. Unexpectedly, the PUA domain was detected also in bacterial and yeast glutamate kinases; this is compatible with the demonstrated role of these enzymes in the regulation of the expression of other genes [1]. It is predicted that the PUA domain is an RNA binding domain.

Number of members: 48

- [1] Aravind L, Koonin EV; Medline: 99193178 "Novel predicted RNA-binding domains associated with the translation machinery." J Mol Evol 1999;48:291-302.

870. (RF1)

eRF1-like proteins

Members of this family are peptide chain release factors. The eukaryotic Release Factor 1 proteins (eRF1s) are involved in termination of translation. The eRF1 protein is functional for all stop codons and appears to abolish read-through of these codons. This family also includes other proteins for which the precise molecular function is unknown. Many of them are from Archaeobacteria. These proteins may also be involved in translation termination but this awaits experimental verification. Number of members: 25

- [1] Frolova L, Le Goff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, et al; Medline: 95082951 "A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor" [see comments] Nature 1994;372:701-703.

- [2] Drugeon G, Jean-Jean O, Frolova L, Le Goff X, Philippe M, Kisselev L, Haenni AL; Medline: 97315314 "Eukaryotic release factor 1 (eRF1) abolishes readthrough and competes with suppressor tRNAs at all three termination codons in messenger RNA." Nucleic Acids Res 1997;25:2254-2258.

731

871. (Ribosomal_L14e)Ribosomal protein L14

This family includes the eukaryotic ribosomal protein L14.

Number of members: 15

5

872. (Ribosomal_S27)

Ribosomal protein S27a

10

This family of ribosomal proteins consists mainly of the 40S ribosomal protein S27a which is synthesized as a C-terminal extension of ubiquitin (CEP). The S27a domain compromises the C-terminal half of the protein. The synthesis of ribosomal proteins as extensions of ubiquitin promotes their incorporation into nascent ribosomes by a transient metabolic stabilization and is required for efficient ribosome biogenesis [3]. The ribosomal extension protein S27a contains a basic region that is proposed to form a zinc finger; its fusion gene is proposed as a mechanism to maintain a fixed ratio between ubiquitin necessary for degrading proteins and ribosomes a source of proteins [2].

15

Number of members: 36

20

873. (Spermine_synth)

Spermine/spermidine synthase

25

Spermine and spermidine are polyamines. This family includes spermidine synthase that catalyses the fifth (last) step in the biosynthesis of spermidine from arginine, and spermine synthase.

Number of members: 39

30

[1] Mezquita J, Pau M, Mezquita C; Medline: 97449308 "Characterization and expression of two chicken cDNAs encoding ubiquitin fused to ribosomal proteins of 52 and 80 amino acids." Gene 1997;195:313-319.

732

[2] Redman KL, Rechsteiner M; Medline: 89181932 "Identification of the long ubiquitin extension as ribosomal protein S27a." Nature 1989;338:438-440.

[3] Finley D, Bartel B, Varshavsky A; Medline: 89181925 "The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis." Nature
5 1989;338:394-401.

874. (Surp)

Surp module

10

[1] Denhez F, Lafyatis R; Medline: 94266805 "Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot." J Biol Chem 1994;269:16170-16179.

15

This domain is also known as the SWAP domain. SWAP stands for Suppressor-of-White-APricot. It has been suggested that these domains may be RNA binding [1].

Number of members: 32

20

875. (TFIIE)

TFIIE alpha subunit

25

The general transcription factor TFIIE has an essential role in eukaryotic transcription initiation together with RNA polymerase II and other general factors. Human TFIIE consists of two subunits TFIIE-alpha Swiss:P29083 and TFIIE-beta Swiss:P29084 and joins the preinitiation complex after RNA polymerase II and TFIIF [1]. This family consists of the conserved amino terminal region of eukaryotic TFIIE-alpha [2] and proteins from archaeobacteria that are presumed to be TFIIE-alpha subunits also Swiss:O29501 [3].

30

Number of members: 12

[1] Ohkuma Y, Sumimoto H, Hoffmann A, Shimasaki S, Horikoshi M, Roeder RG; Medline:

92065982 "Structural motifs and potential sigma homologies in the large subunit of human general transcription factor TFIIE." *Nature* 1991;354:398-401.

[2] Ohkuma Y, Hashimoto S, Roeder RG, Horikoshi M; Medline: 93087200 Identification of two large subdomains in TFIIE-alpha on the basis of homology between *Xenopus* and human sequences. *Nucleic Acids Res* 1992;20:5838-5838.

[3] Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al; Medline: 98049343 "The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon *Archaeoglobus fulgidus*." *Nature* 1997;390:364-370.

876. (Transglut_core)

Cross-reference(s) PS00547; TRANSGLUTAMINASES

Transglutaminases (EC 2.3.2.13) (TGase) [1,2] are calcium-dependent enzymes that catalyze the cross-linking of proteins by promoting the formation of isopeptide bonds between the gamma-carboxyl group of a glutamine in one polypeptide chain and the epsilon-amino group of a lysine in a second polypeptide chain. TGases also catalyze the conjugation of polyamines to proteins. The best known transglutaminase is blood coagulation factor XIII, a plasma tetrameric protein composed of two catalytic A subunits and two non-catalytic B subunits. Factor XIII is responsible for cross-linking fibrin chains, thus stabilizing the fibrin clot. Other forms of transglutaminases are widely distributed in various organs, tissues and body fluids. Sequence data is available for the following forms of TGase:

- Transglutaminase K (Tgase K), a membrane-bound enzyme found in mammalian epidermis and important for the formation of the cornified cell envelope (gene TGM1).
- Tissue transglutaminase (TGase C), a monomeric ubiquitous enzyme located in the cytoplasm (gene TGM2).
- Transglutaminase 3, responsible for the later stages of cell envelope formation in the epidermis and the hair follicle (gene TGM3).
- Transglutaminase 4 (gene TGM4).

A conserved cysteine is known to be involved in the catalytic mechanism of TGases. The erythrocyte membrane band 4.2 protein, which probably plays an important role in regulating the shape of erythrocytes and their mechanical properties, is evolutionary related to TGases.

5 However the active site cysteine is substituted by an alanine and the 4.2 protein does not show TGase activity.

Consensus pattern:[GT]-Q-[CA]-W-V-x-[SA]-[GA]-[IVT]-x(2)-T-x-~~[LMSC]~~[LMSC SEQ ID NO:547]-R-[CSA]-[LV]-G [The first C is the active site residue] Sequences known to
10 belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Ichinose A., Bottenus R.E., Davie E.W. J. Biol. Chem. 265:13411-13414(1990).

[2] Greenberg C.S., Birckbichler P.J., Rice R.H. FASEB J. 5:3071-3077(1991).

877. (TruB_N)

TruB family pseudouridylate synthase (N terminal domain)

20 Members of this family are involved in modifying bases in RNA molecules. They carry out the conversion of uracil bases to pseudouridine. This family includes TruB, a pseudouridylate synthase that specifically converts uracil 55 to pseudouridine in most tRNAs. This family also includes Cbf5p that modifies rRNA [2].

25 Number of members: 33

[1] Nurse K, Wrzesinski J, Bakin A, Lane BG, Ofengand J; Medline: 96079944 "Purification, cloning, and properties of the tRNA psi 55 synthase from Escherichia coli." RNA 1995;1:102-112.

30 [2] Lafontaine DLJ, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D; Medline: 98139521 "The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase." Genes Dev 1998;12:527-537.

878. (UDPGP)

UTP--glucose-1-phosphate uridylyltransferase

5 This family consists of UTP--glucose-1-phosphate uridylyltransferases, EC:2.7.7.9. Also known as UDP-glucose pyrophosphorylase (UDPGP) and Glucose-1-phosphate uridylyltransferase. UTP--glucose-1-phosphate uridylyltransferase catalyses the interconversion of MgUTP + glucose-1-phosphate and UDP-glucose + MgPPi [1]. UDP-glucose is an important intermediate in mammalian carbohydrate interconversion involved in various metabolic roles depending on tissue type [1]. In *Dictyostelium* (slime mold) mutants in this enzyme abort the development cycle [2]. Also within the family is UDP-N-acetylglucosamine Swiss:Q16222 or AGX1 [3] and two hypothetical proteins from *Borrelia burgdorferi* the lyme disease spirochaete Swiss:O51893 and Swiss:O51036.

15 Number of members: 18

[1] Duggleby RG, Chao YC, Huang JG, Peng HL, Chang HY; Medline: 96202932 "Sequence differences between human muscle and liver cDNAs for UDPglucose pyrophosphorylase and kinetic properties of the recombinant enzymes expressed in *Escherichia coli*." *Eur J Biochem* 1996;235:173-179.

[2] Ragheb JA, Dottin RP; Medline: 87231075 "Structure and sequence of a UDP glucose pyrophosphorylase gene of *Dictyostelium discoideum*." *Nucleic Acids Res* 1987;15:3891-3906.

[3] Mio T, Yabe T, Arisawa M, Yamada-Okabe H; Medline: 98269105 "The eukaryotic UDP-N-acetylglucosamine pyrophosphorylases. Gene cloning, protein expression, and catalytic mechanism. *J Biol Chem* 1998;273:14392-14397.

879. (UPF004)

30 Uncharacterized protein family UPF0044 signature

Cross-reference(s) PS01301; UPF0044

The following uncharacterized proteins have been shown [1] to be highly similar:

- *Bacillus subtilis* hypothetical protein yqeI.
- *Escherichia coli* hypothetical protein yhbY and HI1333, the corresponding *Haemophilus influenzae* protein.
- 5 - *Methanococcus jannaschii* hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

10 Consensus pattern: L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)-[LIV]-[GA]-x(2)-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

15 880. (zf-A20)

A20-like zinc finger

A20- (an inhibitor of cell death)-like zinc fingers. The zinc finger mediates self-association in A20. These fingers also mediate IL-1-induced NF-kappa B activation.

20

Number of members: 22

[1] Heyninck K, Beyaert R; Medline: 99126071 "The cytokine-inducible zinc finger protein A20 inhibits IL-1-induced NF- kappaB activation at the level of TRAF6. FEBS Lett
25 1999;442:147-150.

[2] De Valck D, Heyninck K, Van Crielinge W, Contreras R, Beyaert R, Fiers W; Medline: 96390831 "A20, an inhibitor of cell death, self-associates by its zinc finger domain." FEBS Lett 1996;384:61-64.

[3] Song HY, Rothe M, Goeddel DV; Medline: 96270609 "The tumor necrosis factor-inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB
30 activation. Proc Natl Acad Sci U S A 1996;93:6721-6725.

737

[4] Opipari AW Jr, Boguski MS, Dixit VM; Medline: 90368626 "The A20 cDNA induced by tumor necrosis factor alpha encodes a novel type of zinc finger protein." J Biol Chem 1990;265:14705-14708.

5

881. (zf-PARP)

Poly(ADP-ribose) polymerase zinc finger domain

Cross-reference(s) PS00347; PARP_ZN_FINGER_1 PS50064; PARP_ZN_FINGER_2

10

Poly(ADP-ribose) polymerase (EC 2.4.2.30) (PARP) [1,2] is a eukaryotic enzyme that catalyzes the covalent attachment of ADP-ribose units from NAD(+) to various nuclear acceptor proteins. This post-translational modification of nuclear proteins is dependent on DNA. It appears to be involved in the regulation of various important cellular processes such as differentiation, proliferation and tumor transformation as well as in the regulation of the molecular events involved in the recovery of the cell from DNA damage. Structurally, PARP, about 1000 amino-acids residues long, consists of three distinct domains: an N-terminal zinc-dependent DNA-binding domain, a central automodification domain and a C-terminal NAD-binding domain. The DNA-binding region contains a pair of zinc finger domains which have been shown to bind DNA in a zinc-dependent manner. The zinc finger domains of PARP seem to bind specifically to single-stranded DNA. DNA ligase III [3] contains, in its N-terminal section, a single copy of a zinc finger highly similar to those of PARP.

15

20

25

Consensus pattern: C-[KR]-x-C-x(3)-I-x-K-x(3)-[RG]-x(16,18)-W-[FYH]-H-x(2)-C [The three C's and the H are zinc ligands] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT/NONE. Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT/NONE.

30

Note: This documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

[1] Althaus F.R., Richter C.R. Mol. Biol. Biochem. Biophys. 37:1-126(1987).

[2] de Murcia G., Menissier de Murcia J. Trends Biochem. Sci. 19:172-176(1994).

[3] Wei Y.-F., Robins P., Carter K., Caldecott K., Pappin D.J.C., Yu G.-L., Wang R.-P.,
5 Shell B.K., Nash R.A., Schar P., Barnes D.E., Haseltine W.A., Lindahl T. Mol. Cell. Biol.
15:3206-3216(1995).

882. Adenylylsulfate kinase (APS_kinase)

Enzyme that catalyses the phosphorylation of adenylylsulfate to 3'-phosphoadenylylsulfate.

10 This domain contains an ATP binding P-loop motif. Number of members: 34

[1] MacRae IJ, Rose AB, Segel IH; Medline: 99003196 "Adenosine 5'-phosphosulfate kinase
from *Penicillium chrysogenum*. site- directed mutagenesis at putative phosphoryl-accepting
and ATP P-loop residues. J Biol Chem 1998;273:28583-28589.

15

883. DNA polymerase family B signature DNA_POLYMERASE_B (DNA_pol_B)

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the
accurate replication of DNA. They require either a small RNA molecule or a protein as a
20 primer for the de novo synthesis of a DNA chain. On the basis of sequence similarity, a
number of DNA polymerases have been grouped [1 to 7] under the designation of DNA
polymerase family B. These are:

- Higher eukaryotes polymerases alpha.
- Higher eukaryotes polymerases delta.
- 25 - Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase
III/delta (gene POL3) and polymerase REV3.
- *Escherichia coli* polymerase II (gene *dinA* or *polB*).
- Archaeobacterial polymerases.
- Polymerases of viruses from the herpesviridae family.
- 30 - Polymerases from Adenoviruses.
- Polymerases from Baculoviruses.
- Polymerases from *Chlorella* viruses.
- Polymerases from Poxviruses.

- Bacteriophage T4 polymerase.
 - Podoviridae bacteriophages Phi-29, M2 and PZA polymerase.
 - Tectiviridae bacteriophage PRD1 polymerase.
 - Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants
- 5 (Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bitorquis pEM, Ascobolus immersus pAI2, Claviceps purpurea pCLK1, Neurospora Kalilo and Maranhar, maize S-1, etc).

Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a conserved tetrapeptide with two

10 aspartate residues. Its function is not yet known. However, it has been suggested [3] that it may be involved in binding a magnesium ion. This conserved region was selected as a signature for this family of DNA polymerases.

Consensus pattern [YA]-[GLIVMSTAC][GLIVMSTAC SEQ ID NO:723]-D-T-D-[SG]-

15 [LIVMFTC][LIVMFTC SEQ ID NO:724])-x-[LIVMSTAC][LIVMSTAC SEQ ID NO:151]]

Sequences known to belong to this class detected by the patternALL, except for yeast polymerase II/epsilon, Agaricus bitorquis pEM and Sulfolobus solfataricus polymerase II.

[1] Jung G., Leavitt M.C., Hsieh J.-C., Ito J. Proc. Natl. Acad. Sci. U.S.A. 84:8287-8291(1987).

20

[2] Bernad A., Zaballos A., Salas M., Blanco L. EMBO J. 6:4219-4225(1987).

[3] Argos P. Nucleic Acids Res. 16:9909-9916(1988).

[4] Wang T.S.-F., Wong S.W., Korn D. FASEB J. 3:14-21(1989).

[5] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461-467(1990).

25 [6] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[7] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787-802(1993).

884. DNA polymerase family X signature - DNA_POLYMERASE_X (DNA_polymeraseX)

30

DNA polymerases (EC 2.7.7.7) can be classified, on the basis of sequence similarity [1], into at least four different groups: A, B, C and X. DNA polymerases that belong to family X are listed below [2]:

- Vertebrate polymerase beta, involved in DNA repair.
- Yeast polymerase IV (POL4) [3], an enzyme with similar characteristics to that of the mammalian polymerase beta.
- Terminal deoxynucleotidyltransferase (TdT) (EC 2.7.7.31). TdT catalyzes the elongation of polydeoxynucleotide chains by terminal addition. One of the functions of this enzyme is the addition of nucleotides at the junction of rearranged Ig heavy chain and T cell receptor gene segments during the maturation of B and T cells.
- African Swine Fever virus protein O174L [4].
- Fission yeast hypothetical protein SpAC2F7.06c.

These enzymes are small (about 40 Kd) compared with other polymerases and their reaction mechanism operates via a distributive mode, i.e. they dissociate from the template-primer after addition of each nucleotide.

As a signature pattern for this family of DNA polymerases, a highly conserved region that contains a conserved arginine and two conserved aspartic acid residues were selected. The latter together with the arginine have been shown [5] to be involved in primer binding in polymerase beta.

Consensus pattern G-[SG]-[LFY]-x-R-[GE]-x(3)-[SGCL][SGCL SEQ ID NO:725]-x-D-[LIVM][LIVM SEQ ID NO:4]-D-[LIVMFY][LIVMFY SEQ ID NO:18](3)-x(2)-[SAP]
Sequences known to belong to this class detected by the patternALL.

[1] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[2] Matsukage A., Nishikawa K., Ooi T., Seto Y., Yamaguchi M. J. Biol. Chem. 262:8960-8962(1987).

[3] Prasad R., Widen S.G., Singhal R.K., Watkins J., Prakash L., Wilson S.H. Nucleic Acids Res. 21:5301-5307(1993).

[4] Yanez R.J., Rodriguez J.M., Nogal M.L., Yuste L., Enriquez C., Rodriguez J.F., Vinuela E. Virology 208:249-278(1995).

[5] Date T., Yamamoto S., Tanihara K., Nishimoto Y., Matsukage A. Biochemistry 30:5286-5292(1991).

885. DUF14 - Domain of unknown function

This domain is found in glutamate synthase, tungsten formylmethanofuran dehydrogenase subunit c (FwdC) and molybdenum formylmethanofuran dehydrogenase subunit c (FmdC). It has no known function. Number of members: 52

5

[1] Hochheimer A, Hedderich R, Thauer RK; Medline: 99035764. "The formylmethanofuran dehydrogenase isoenzymes in *Methanobacterium wolfei* and *Methanobacterium thermoautotrophicum*: induction of the molybdenum isoenzyme by molybdate and constitutive synthesis of the tungsten isoenzyme." Arch Microbiol 1998;170:389-393.

10

886. DUF18-Domain of unknown function

This domain of unknown function is found in several *C. elegans* proteins. The domain is 120 amino acids long and rich in cysteine residues. There are 16 conserved cysteine positions in the domain. Number of members: 34

15

887. DUF27-Domain of unknown function

This domain is found in a number of otherwise unrelated proteins. This domain is found at the C-terminus of the macro-H2A histone protein Swiss:Q02874. This domain is found in the non-structural proteins of several types of ssRNA viruses such as NSP2 from alphaviruses Swiss:P03317. This domain is also found on its own in a family of proteins from bacteria Swiss:P75918, archaeobacteria Swiss:O59182 and eukaryotes Swiss:Q17432, suggesting that it is involved in an important and ubiquitous cellular process. Number of members: 66

20

888. DUF37-Domain of unknown function

This domain is found in short (70 amino acid) hypothetical proteins from various bacteria. The domain contains three conserved cysteine residues. Swiss:Q44066 from *Aeromonas hydrophila* has been found to have hemolytic activity (unpublished). Number of members: 19

25

889. EGF-like domain signatures. (EGF-like)

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less

30

conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Adipocyte differentiation inhibitor (gene PREF-1) from mouse (6 copies).
- Agrin, a basal lamina protein that causes the aggregation of acetylcholine receptors on cultured muscle fibers (4 copies).
- Amphiregulin, a growth factor (1 copy).
- Betacellulin, a growth factor (1 copy).
- Blastula proteins BP10 and Span from sea urchin which are thought to be involved in pattern formation (1 copy).
- BM86, a glycoprotein antigen of cattle tick (7 copies).
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity (1-2 copies). Homologous proteins are found in sea urchin - suBMP (1 copy) - and in Drosophila - the dorsal-ventral patterning protein tolloid (2 copies).
- Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Caenorhabditis elegans APX-1 protein, a patterning protein (4.5 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type I and IV collagen and fibronectin (1 copy).
- Cartilage matrix protein CMP (1 copy).
- Cartilage oligomeric matrix protein COMP (4 copies).
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit ASGP-2 from rat (2 copies).
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- Complement C1r components (1 copy).
- Complement C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Crumbs, an epithelial development protein from Drosophila (29 copies).
- Epidermal growth factor precursor (7-9 copies).
- Exogastrula-inducing peptides A, C, D and X from sea urchin (1 copy).
- Fat protein, a Drosophila cadherin-related tumor suppressor (5 copies).

- Fetal antigen 1, a probable neuroendocrine differentiation protein, which is derived from the delta-like protein (DLK) (6 copies).
- Fibrillin 1 (47 copies) and fibrillin 2 (14 copies).
- Fibropellins IA (21 copies), IB (13 copies), IC (8 copies), II (4 copies) and III (8 copies)
- 5 from the apical lamina - a component of the extracellular matrix - of sea urchin.
- Fibulin-1 and -2, two extracellular matrix proteins (9-11 copies).
- Giant-lens protein (protein Argos), which regulates cell determination and axon guidance in the *Drosophila* eye (1 copy).
- Growth factor-related proteins from various poxviruses (1 copy).
- 10 - Gurken protein, a *Drosophila* developmental protein (1 copy).
- Heparin-binding EGF-like growth factor (HB-EGF), transforming growth factor alpha (TGF-alpha), growth factors Lin-3 and Spitz (1 copy); the precursors are membrane proteins, the mature form is located extracellular.
- Hepatocyte growth factor (HGF) activator (EC 3.4.21.-) (2 copies).
- 15 - LDL and VLDL receptors, which bind and transport low-density lipoproteins and very low-density lipoproteins (3 copies).
- LDL receptor-related protein (LRP), which may act as a receptor for endocytosis of extracellular ligands (22 copies).
- Leucocyte antigen CD97 (3 copies), cell surface glycoprotein EMR1 (6 copies) and cell
- 20 surface glycoprotein F4/80 (7 copies).
- Limulus clotting factor C, which is involved in hemostasis and host defense mechanisms in Japanese horseshoe crab (1 copy).
- Meprin A alpha subunit, a mammalian membrane-bound endopeptidase (1 copy).
- Milk fat globule-EGF factor 8 (MFG-E8) from mouse (2 copies).
- 25 - Neuregulin GGF-I and GGF-II, two human glial growth factors (1 copy).
- Neurexins from mammals (3 copies).
- Neurogenic proteins Notch, Xotch and the human homolog Tan-1 (36 copies), Delta (9 copies) and the similar differentiation proteins Lag-2 from *Caenorhabditis elegans* (2 copies), Serrate (14 copies) and Slit (7 copies) from *Drosophila*.
- 30 - Nidogen (also called entactin), a basement membrane protein from chordates (2-6 copies).
- Ookinete surface proteins (24 Kd, 25 Kd, 28 Kd) from *Plasmodium* (4 copies).
- Pancreatic secretory granule membrane major glycoprotein GP2 (1 copy).
- Perforin, which lyses non-specifically a variety of target cells (1 copy).

- Proteoglycans aggrecan (1 copy), versican (2 copies), perlecan (at least 2 copies), brevican (1 copy) and chondroitin sulfate proteoglycan (gene PG-M) (2 copies).
- Prostaglandin G/H synthase 1 and 2 (EC 1.14.99.1) (1 copy), which is found in the endoplasmatic reticulum.
- 5 - S1-5, a human extracellular protein whose ultimate activity is probably modulated by the environment (5 copies).
- Schwannoma-derived growth factor (SDGF), an autocrine growth factor as well as a mitogen for different target cells (1 copy).
- Selectins. Cell adhesion proteins such as ELAM-1 (E-selectin), GMP-140 (P-selectin), or
10 the lymph-node homing receptor (L-selectin) (1 copy).
- Serine/threonine-protein kinase homolog (gene Pro25) from *Arabidopsis thaliana*, which may be involved in assembly or regulation of light-harvesting chlorophyll A/B protein (2 copies).
- Sperm-egg fusion proteins PH-30 alpha and beta from guinea pig (1 copy).
- 15 - Stromal cell derived protein-1 (SCP-1) from mouse (6 copies).
- TDGF-1, human teratocarcinoma-derived growth factor 1 (1 copy).
- Tenascin (or neuronectin), an extracellular matrix protein from mammals (14.5 copies), chicken (TEN-A) (13.5 copies) and the related proteins human tenascin-X (18 copies) and tenascin-like proteins TEN-A and TEN-M from *Drosophila* (8 copies).
- 20 - Thrombomodulin (fetomodulin), which together with thrombin activates protein C (6 copies).
- Thrombospondin 1, 2 (3 copies), 3 and 4 (4 copies), adhesive glycoproteins that mediate cell-to-cell and cell-to-matrix interactions.
- Thyroid peroxidase 1 and 2 (EC 1.11.1.8) from human (1 copy).
- 25 - Transforming growth factor beta-1 binding protein (TGF-B1-BP) (16 or 18 copies).
- Tyrosine-protein kinase receptors Tek and Tie (EC 2.7.1.112) (3 copies).
- Urokinase-type plasminogen activator (EC 3.4.21.73) (UPA) and tissue plasminogen activator (EC 3.4.21.68) (TPA) (1 copy).
- Uromodulin (Tamm-horsfall urinary glycoprotein) (THP) (3 copies).
- 30 - Vitamin K-dependent anticoagulants protein C (2 copies) and protein S (4 copies) and the similar protein Z, a single-chain plasma glycoprotein of unknown function (2 copies).
- 63 Kd sperm flagellar membrane protein from sea urchin (3 copies).
- 93 Kd protein (gene nel) from chicken (5 copies).

- Hypothetical 337.6 Kd protein T20G5.3 from *Caenorhabditis elegans* (44 copies).

The functional significance of EGF domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted (exception: prostaglandin G/H synthase). The EGF domain includes six cysteine residues which have been shown (in EGF) to be involved in disulfide bonds. The main structure is a two-stranded beta-sheet followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines strongly vary in length as shown in the following schematic representation of the EGF-like domain:

```

      +-----+      +-----+      |      |      |
| x(4)-C-x(0,48)-C-x(3,12)-C-x(1,70)-C-x(1,6)-C-x(2)-G-a-x(0,21)-G-x(2)-C-x  |
| *****
      +-----+

```

'C': conserved cysteine involved in a disulfide bond.

'G': often conserved glycine

'a': often conserved aromatic amino acid

'*': position of both patterns.

'x': any residue

The region between the 5th and 6th cysteine contains two conserved glycines of which at least one is present in most EGF-like domains. Two patterns were created for this domain, each including one of these C-terminal conserved glycine residues.

Consensus pattern: C-x-C-x(5)-G-x(2)-C [The 3 C's are involved in disulfide bonds]

Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT87 proteins, of which 27 can be considered as possible candidates.

Consensus pattern: C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C [The three C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern A majority, but not

those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT83 proteins, of which 49 can be considered as possible candidates. Note The beta chain of the integrin family of proteins contains 2 cysteine- rich repeats which were said to be dissimilar with the EGF pattern [7].

Note Laminin EGF-like repeats (see <PDOC00961>) are longer than the average EGF module and contain a further disulfide bond C-terminal of the EGF-like region. Perlecan and agrin contain both EGF-like domains and laminin-type EGF-like domains. Note the pattern do not detect all of the repeats of proteins with multiple EGF-like repeats. Note see <PDOC00913> for an entry describing specifically the subset of EGF- like domains that bind calcium.

[1] Davis C.G. New Biol. 2:410-419(1990).

[2] Blomquist M.C., Hunt L.T., Barker W.C. Proc. Natl. Acad. Sci. U.S.A. 81:7363-7367(1984).

[3] Barker W.C., Johnson G.C., Hunt L.T., George D.G. Protein Nucl. Acid Enz. 29:54-68(1986).

[4] Doolittle R.F., Feng D.F., Johnson M.S. Nature 307:558-560(1984).

[5] Appella E., Weber I.T., Blasi F. FEBS Lett. 231:1-4(1988).

[6] Campbell I.D., Bork P. Curr. Opin. Struct. Biol. 3:385-392(1993).

[7] Tamkun J.W., DeSimone D.W., Fonda D., Patel R.S., Buck C., Horwitz A.F., Hynes R.O. Cell 46:271-282(1986).

890. Ham1 family (Ham1p_like)

This family consists of the HAM1 protein Swiss:P47119 and hypothetical archaeal bacterial and C. elegans proteins. HAM1 controls 6-N-hydroxylaminopurine (HAP) sensitivity and mutagenesis in S. cerevisiae Swiss:P47119 [1]. The HAM1 protein protects the cell from HAP, either on the level of deoxynucleoside triphosphate or the DNA level by a yet unidentified set of reactions [1]. Number of members: 19

[1] Noskov VN, Staak K, Shcherbakova PV, Kozmin SG, Negishi K, Ono BC, Hayatsu H, Pavlov YI; Medline: 96381244 "HAM1, the gene controlling 6-N-hydroxylaminopurine sensitivity and mutagenesis in the yeast *Saccharomyces cerevisiae*." Yeast 1996;12:17-29.

5

891. (HCO₃⁻_cotransp)

Anion exchange is a cellular transport function which contributes to the regulation of cell pH and volume. Anion exchangers are a family of functionally related proteins that contributes to these properties by maintaining the intracellular level of the two principal anions: chloride and HCO₃⁻. The best characterized anion exchanger is the band 3 protein [1], which is an erythrocyte anion exchange membrane glycoprotein. Band 3 is a protein of about 900 amino acids which consists of a cytoplasmic N-terminal domain of about 400 residues and an hydrophobic C-terminal section of about 500 residues that contains at least ten transmembrane regions. The cytoplasmic domain provides binding sites for cytoskeletal proteins, while the integral membrane domain is responsible for anion transport. Band 3 protein is specific to erythroid cells, at least two other proteins [2] structurally and functionally related to band 3, are found in nonerythroid tissues:

- AE2 (or B3 related protein; B3RP), a protein of 1200 residues, which seems to be present in a variety of cell types including lymphoid, kidney, and choroid plexus.
- AE3, a protein of 1200 residues, which is specific to neurons.

Structurally AE2 and AE3 are very similar to band 3, the main difference being an extension of some 300 residues of the N-terminal domain in AE2 and AE3.

Two signature patterns were developed for these proteins. The first pattern is based on a conserved stretch of sequence that contains four clustered positive charged residues and which is located at the C-terminal extremity of the cytoplasmic domain, just before the first transmembrane segment from the integral domain. The second pattern is based on the perfectly conserved sequence of the fifth transmembrane segment; this segment contains a lysine, which is the covalent binding site for the isothiocyanate group of DIDS, an inhibitor of anion exchange.

30

Consensus pattern F-G-G-[LIVM][LIVM SEQ ID NO:4])(2)-[KR]-D-[LIVM][LIVM SEQ ID NO:4)]-[RK]-R-R-Y Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FI]-L-I-S-L-I-F-I-Y-E-T-F-x-K-L Sequences known to belong to this class detected by the pattern ALL.

- 5 [1] Jay D., Cantley L. Annu. Rev. Biochem. 55:511-538(1986).
[2] Reithmeier R.A.F. Curr. Opin. Struct. Biol. 3:515-523(1993).

892. ATP phosphoribosyltransferase signature (HisG)

- 10 ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern a region located in the C-terminal part of this enzyme was selected.

Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]

- 15 Sequences known to belong to this class detected by the pattern ALL.

893. HNH endonuclease (HNH)

Number of members: 56

- 20 [1] Shub DA, Goodrich-Blair H, Eddy SR; Medline: 95117127 "Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns." Trends Biochem Sci 1994;19:402-404.
[2] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site- specific endonucleases
25 and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.
[3] Gorbalenya AE; Medline: 95004046 "Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family." Protein Sci 1994;3:1117-1120.

30 894. NEUROHYPOPHYS_HORM (hormone5)

Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophysial peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct

antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels. Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, 5 glumitocin, aspartocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (annetocin), octopi (cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs (conopressins G and S) [2]. The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.

10 Consensus pattern C-~~LIFY~~[LIFY SEQ ID NO:580](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide bond]. Sequences known to belong to this class detected by the pattern ALL.

[1] Acher R., Chauvet J. Biochimie 70:1197-1207(1988).

15 [2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B.T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995).

895. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)

20 All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides. 7,8-dihydro-6-hydroxymethylpterin-
25 pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate. Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote *Pneumocystis carinii*, HPPK is the central domain of a multifunctional folate
30 synthesis enzyme (gene fas) [2]. As a signature for HPPK, a conserved region located in the central section of these enzymes was selected.

750

Consensus pattern [KRHD][KRHD SEQ ID NO:726]-x-[GA]-[PSAE][PSAE SEQ ID NO:727]-R-x(2)-D-[LIV]-D-[LIVM][LIVM SEQ ID NO:4](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

5

[1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S. J. Bacteriol. 174:5971-5977(1992).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

10

896. Metalloenzyme superfamily (Metalloenzyme)

This family includes phosphopentomutase Swiss:P07651 and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, Swiss:P37689. This family is also related to

15 alk_phosphatase [1]. The alignment contains the most conserved residues that are probably involved in metal binding and catalysis. Number of members: 34

[1] Galperin MY, Bairoch A, Koonin EV; Medline: 99180418 "A superfamily of metalloenzymes unifies phosphopentomutase and cofactor- independent phosphoglycerate mutase with alkaline phosphatases and sulfatases." Protein Sci 1998;7:1829-1835.

20

897. Penicillin amidase (Penicil_amidase)

Penicillin amidase or penicillin acylase EC:3.5.1.11 catalyses the hydrolysis of

25 benzylpenicillin to phenylacetic acid and 6-aminopenicillanic acid (6-APA) a key intermediate in the the synthesis of penicillins [1]. Also in the family is cephalosporin acylase Swiss:P07662 and Swiss:P29958 aculeacin A acylase which are involved in the synthesis of related peptide antibiotics. Number of members: 13

30

[1] Verhaert RM, Riemens AM, van der Laan JM, van Duin J, Quax WJ; Medline: 97438505 "Molecular cloning and analysis of the gene encoding the thermostable penicillin G acylase from *Alcaligenes faecalis*. Appl Environ Microbiol 1997;63:3412-3418.

[2] Duggleby HJ, Tolley SP, Hill CP, Dodson EJ, Dodson G, Moody PC; Medline: 95115804
“Penicillin acylase has a single-amino-acid catalytic centre.” *Nature* 1995;373:264-268.

5 898. Phosphoribosyl-AMP cyclohydrolase (PRA-CH)

This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions for activity. Number of members: 13

10 [1] D'Ordine RL, Klem TJ, Davisson VJ; Medline: 99129952 “N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme. *Biochemistry* 1999;38:1537-1546.

899. Phosphoribosyl-ATP pyrophosphohydrolase (PRA-PH)

15 This enzyme catalyses the second step in the histidine biosynthetic pathway. Number of members: 32

20 [1] Keesey JK Jr, Bigelis R, Fink GR; Medline: 79216449 “The product of the *his4* gene cluster in *Saccharomyces cerevisiae*. A trifunctional polypeptide.” *J Biol Chem* 1979 Aug 10;254:7427-7433.

[2] Bruni CB, Carlomagno MS, Formisano S, Paoletta G; Medline: 86310274 “Primary and secondary structural homologies between the *HIS4* gene product of *Saccharomyces cerevisiae* and the *hisIE* and *hisD* gene products of *Escherichia coli* and *Salmonella typhimurium*.” *Mol Gen Genet* 1986;203:389-396.

25

900. Prokaryotic membrane lipoprotein lipid attachment site (PstS)

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase
30 recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene *lpp*).

- *Escherichia coli* lipoprotein-28 (gene nlpA).
- *Escherichia coli* lipoprotein-34 (gene nlpB).
- *Escherichia coli* lipoprotein nlpC.
- *Escherichia coli* lipoprotein nlpD.
- 5 - *Escherichia coli* osmotically inducible lipoprotein B (gene osmB).
- *Escherichia coli* osmotically inducible lipoprotein E (gene osmE).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene pal).
- *Escherichia coli* rare lipoproteins A and B (genes rplA and rplB).
- *Escherichia coli* copper homeostasis protein cutF (or nlpE).
- 10 - *Escherichia coli* plasmids traT proteins.
- *Escherichia coli* Col plasmids lysis proteins.
- A number of *Bacillus* beta-lactamases.
- *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA).
- *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB).
- 15 - *Borrelia hermsii* variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- *Chlamydia trachomatis* outer membrane protein 3 (gene omp3).
- *Fibrobacter succinogenes* endoglucanase cel-3.
- *Haemophilus influenzae* proteins Pal and Pcp.
- *Klebsiella* pullulunase (gene pulA).
- 20 - *Klebsiella* pullulunase secretion protein pulS.
- *Mycoplasma hyorhinis* protein p37.
- *Mycoplasma hyorhinis* variant surface antigens A, B, and C (genes vlpABC).
- *Neisseria* outer membrane protein H.8.
- *Pseudomonas aeruginosa* lipopeptide (gene lppL).
- 25 - *Pseudomonas solanacearum* endoglucanase egl.
- *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
- *Rickettsia* 17 Kd antigen.
- *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
- *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
- 30 - *Treponema pallidum* 34 Kd antigen.
- *Treponema pallidum* membrane protein A (gene tmpA).
- *Vibrio harveyi* chitobiase (gene chb).
- *Yersinia* virulence plasmid protein yscJ.

753

- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion). From the precursor sequences of all these proteins, a consensus pattern was derived and a set of rules to identify this type of post-translational modification.

5

Consensus pattern ~~{DERK}~~{DERK SEQ ID NO:354}}(6)-
~~[LIVMFWSTAG]~~[LIVMFWSTAG SEQ ID NO:352}}(2)-
~~[LIVMFYSTAGCQ]~~[LIVMFYSTAGCQ SEQ ID NO:353}}-[AGS]-C [C is the lipid
 attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the
 sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven
 positions of the sequence. Sequences known to belong to this class detected by the
 patternALL. Other sequence(s) detected in SWISS-PROTsome 100 prokaryotic proteins.
 Some of them are not membrane lipoproteins, but at least half of them could be.

10

15

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).
 [2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).
 [3] von Heijne G. Protein Eng. 2:531-534(1989).
 [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol.
 Chem. 269:14939-14945(1994).

20

901. Ribosome recycling factor (RRF)

The ribosome recycling factor (RRF / ribosome release factor) dissociates the ribosome from the mRNA after termination of translation, and is essential bacterial growth [1]. Thus
 ribosomes are "recycled" and ready for another round of protein synthesis. Number of
 members: 27

25

[1] Janosi L, Shimizu I, Kaji A; Medline: 94240115 "Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth." Proc Natl Acad Sci U S A 1994;91:4249-
 4253.

30

902. S-layer homology(SLH)

S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the

peptidoglycan [3]. The SLH domain has been found in:

- S-layer glycoprotein of *Acetogenium kivui* (3 copies).
- S-layer 125 Kd protein of *Bacillus sphaericus* (3 copies).
- S-layer protein of *Bacillus anthracis* (3 copies).
- S-layer protein of *Bacillus licheniformis* (3 copies).
- S-layer protein (HWP) from *Bacillus brevis* strain HPD31 (3 copies).
- Middle cell wall protein (MWP) from *Bacillus brevis* strain 47 (3 copies).
- S-layer protein (p100) of *Thermus thermophilus* (1 copy).
- Outer membrane protein Omp-alpha from *Thermotoga maritima* (1 copy).
- Cellulosome anchoring protein (gene *ancA*), outer layer protein B (OlpB) and a further potential cell surface glycoprotein from *Clostridium thermocellum* (3 copies; the first copy is missing its N-terminal third which is appended to the end of the third copy; may have arisen by circular permutation).
- Amylopullulanase (gene *amyB*) from *Thermoanaerobacter thermosulfurogenes* (3 copies)
- Amylopullulanase (gene *aapT*) from *Bacillus* strain XAL-601 (3 copies).
- Endoglucanase from *Bacillus* strain KSM-635 (3 copies).
- Exoglucanase (gene *xynX*) from *Clostridium thermocellum* (3 copies).
- Xylanase A (gene *xynA*) from *Thermoanaerobacter saccharolyticum* (2 copies; 3 copies if a frameshift is taken into account).
- Protein involved in butirosin production (ButB) from *Bacillus circulans* (2 incomplete copies; 3 copies if three frameshifts are taken into account).
- Two hypothetical proteins from *Synechocystis* strain PCC 6803 (1 copy each).
- A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase gene from *Bacillus circulans* (fragment of 1 copy; 3 copies if two frameshifts are taken into account).

SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH sequences are fairly divergent with an average identity of about 25%. It is however possible

to build a sequence pattern that starts at the second position of the domain and that spans 3/4 of its length.

Consensus pattern[LVFYF][LVFYT SEQ ID NO:728)]-x-[DA]-x(2,5)-

[DNGSATPHY][DNGSATPHY SEQ ID NO:729)]-[FYWPDA][FYWPDA SEQ ID NO:730)]-x(4)-[LIV]-x(2)-[GTALV][GTALV SEQ ID NO:731)]-x(4,6)-

[LIVFYC][LIVFYC SEQ ID NO:732)]-x(2)-G-x-[PGSTA][PGSTA SEQ ID NO:733)]-

x(2,3)-[MFYA][MFYA SEQ ID NO:734)]-x-[PGAV][PGAV SEQ ID NO:735)]-x(3,10)-

[LIVMA][LIVMA SEQ ID NO:30)]-[STKR][STKR SEQ ID NO:152)]-[RY]-x-[EQ]-x-

[STALIVM][STALIVM SEQ ID NO:736)] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Beveridge T.J. Curr. Opin. Struct. Biol. 4:204-212(1994).

[2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. J. Bacteriol. 176:1224-1233(1994).

[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. J. Bacteriol. 177:2451-2459(1995).

903. Queuine tRNA-ribosyltransferase (TGT)

This is a family of queuine tRNA-ribosyltransferases EC:2.4.2.29, also known as tRNA-guanine transglycosylase and guanine insertion enzyme. Queuine tRNA-ribosyltransferase modifies tRNAs for asparagine, aspartic acid, histidine and tyrosine with queuine. It catalyses the exchange of guanine-34 at the wobble position with 7-aminomethyl-7-deazaguanine, and the addition of a cyclopentenediol moiety to 7-aminomethyl-7-deazaguanine-34 tRNA; giving a hypermodified base queuine in the wobble position [1,2]. The aligned region contains a zinc binding motif C-x-C-x2-C-x29-H, and important tRNA and 7-aminomethyl-7deazaguanine binding residues [1]. Number of members: 27

[1] Romier C, Reuter K, Suck D, Ficner R; Medline: 96256303 "Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange." EMBO J 1996;15:2850-2857.

[2] Garcia GA, Koch KA, Chong S; Medline: 93287116 "tRNA-guanine transglycosylase from *Escherichia coli*. Overexpression, purification and quaternary structure." *J Mol Biol* 1993;231:489-497.

5

904. ThiC Family (ThiC)

ThiC is found within the thiamine biosynthesis operon. ThiC is involved in pyrimidine biosynthesis [2]. ThiC catalyzes the substitution of the pyrophosphate of 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine phosphate [3]. Number of members: 12

10

[1] Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Medline: 93163063 "Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in *Escherichia coli* K-12." *J Bacteriol* 1993;175:982-992.

15

[2] Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi J; Medline: 99311269 "Thiamin biosynthesis in prokaryotes." *Arch Microbiol* 1999;171:293-300.

[3] Zhang Y, Taylor SV, Chiu HJ, Begley TP; Medline: 97284509 "Characterization of the *Bacillus subtilis* thiC operon involved in thiamine biosynthesis." *J Bacteriol* 1997;179:3030-3035.

20

905. Putative tRNA binding domain (tRNA_bind)

This domain is found in prokaryotic methionyl-tRNA synthetases, prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], and endothelial-monocyte activating polypeptide II. G4p1 binds specifically to tRNA form a complex with methionyl-tRNA synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct tRNA to the active site of the enzyme [2]. This domain may perform a

25

30

common function in tRNA aminoacylation [1]. Number of members: 12

[1] Kleeman TA, Wei D, Simpson KL, First EA; Medline: 97306356 "Human tyrosyl-tRNA synthetase shares amino acid sequence homology with a putative cytokine." J Biol Chem 1997;272:14420-14425.

[2] Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Mann M, Hurt EC; Medline: 97050848 "The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases." EMBO J 1996;15:5437-5448.

906. UbiA prenyltransferase family signature (UbiA)

The following prenyltransferases are evolutionary related [1,2]:

- Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA).
- Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2).
- Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene COX10) and from bacteria (genes cyoE or ctaB).

These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and was used as a signature pattern.

Consensus pattern N-x(3)-[DE]-x(2)-[LIF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT/NONE.

[1] Melzer M., Heide L. Biochim. Biophys. Acta 1212:93-102(1994).

[2] Mogi T., Saiki K., Anraku Y. Mol. Microbiol. 14:391-398(1994).

907. Uncharacterized protein family UPF0044 signature (UPF0044)

The following uncharacterized proteins have been shown [1] to be highly similar:

- Bacillus subtilis hypothetical protein yqel.
- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.
- Methanococcus jannaschii hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)-[LIV]-
5 [GA]-x(2)-G Sequences known to belong to this class detected by the patternALL.

908. ATP synthase (C/AC39) subunit (vATP-synt_AC39)

10 This family includes the AC39 subunit from vacuolar ATP synthase Swiss:P32366 [1], and the C subunit from archaeobacterial ATP synthase [2]. The family also includes subunit C from the Sodium transporting ATP synthase from *Enterococcus hirae* Swiss:P43456 [3].
Number of members: 12

15 [1] Bauerle C, Ho MN, LinJorfer MA, Stevens TH; Medline: 93286119 "The *Saccharomyces cerevisiae* VMA6 gene encodes the 36-kDa subunit of the vacuolar H(+)-ATPase membrane sector." J Biol Chem 1993;268:12749-12757.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." J Biol Chem 1996;271:18843-18852.

20 [3] Takase K, Kakinuma S, Yamato I, Konishi K, Igarashi K, Kakinuma Y; Medline: 94209269 "Sequencing and characterization of the ntp gene cluster for vacuolar- type Na(+)-translocating ATPase of *Enterococcus hirae*." J Biol Chem 1994;269:11037-11044.

25 909. ATP synthase (E/31 kDa) subunit (vATP-synt_E)

This family includes the vacuolar ATP synthase E subunit [1], as well as the archaeobacterial ATP synthase E subunit [2]. Number of members: 24

30 [1] Foury F; Medline: 91009356 "The 31-kDa polypeptide is an essential subunit of the vacuolar ATPase in *Saccharomyces cerevisiae*." J Biol Chem 1990;265:18554-18560.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." J Biol Chem 1996;271:18843-18852.

910. (WW)

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline- motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C- terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

- Utrophin, a dystrophin-like protein of unknown function.

- Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

- Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

- Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>, followed by a histidine-rich region, 3 WW domains and a HECT domain.

- Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

5 - Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

- Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

10 - IQGAP, a human GTPase activating protein acting on ras. It contains an N- terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

- Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2- type myosin, each containing two WW-domains at the N-terminus.

15 - Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

- Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

20 Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE][GSTNE SEQ ID NO:737]-[GSTQCR][GSTQCR SEQ ID NO:738]-[FYW]-x(2)-P Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT8. Sequences known to belong to this class detected by the profileALL.

25 [1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

[5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

30 [6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

911. Xeroderma pigmentosum (XP) [1] (XPG_1)

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair.

- 5 There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2].

XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets:

- 10 - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3'incision in human DNA nucleotide excision repair [9].
- Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease.

15

In addition to the proteins listed in the above groups, this family also includes:

- Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs.
- Yeast EXO1 (DHS1), a protein with probably the same function as exo1.
20 - Yeast DIN7.

Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the
25 C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset.

30

Two signature patterns were developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide.

Consensus pattern [VI]-[KRE]-P-x-~~[FYIL]~~[FYIL SEQ ID NO:644]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

5

Consensus pattern [GS]-~~[LIVM]~~[LIVM SEQ ID NO:4]-[PER]-[FYS]-~~[LIVM]~~[LIVM SEQ ID NO:4]-x-A-P-x-E-A-[DE]-[PAS]-[QS]-[CLM] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

- 10 [1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).
 [2] Scherly D., Nospikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).
 [3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).
 15 [4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).
 [5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).
 [6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).
 [7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).
 20 [8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).
 [9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

25

912. 5-formyltetrahydrofolate cyclo-ligase (5-FTHF_cyc-lig)

5-formyltetrahydrofolate cyclo-ligase or methenyl-THF synthetase EC:6.3.3.2 catalyses the interchange of 5-formyltetrahydrofolate (5-FTHF) to 5-10-methenyltetrahydrofolate, this
 30 requires ATP and Mg²⁺ [1]. 5-FTHF is used in chemotherapy where it is clinically known as Leucovorin [2].

Number of members: 23

763

[1] Dayan A, Bertrand R, Beauchemin M, Chahla D, Mamo A, Filion M, Skup D, Massie B, Jolivet J; Medline: 96096540 "Cloning and characterization of the human 5,10-methenyltetrahydrofolate synthetase-encoding cDNA." Gene 1995;165:307-311.

[2] Maras B, Stover P, Valiante S, Barra D, Schirch V; Medline: 94308074 "Primary structure and tetrahydropteroylglutamate binding site of rabbit liver cytosolic 5,10-methenyltetrahydrofolate synthetase." J Biol Chem 1994;269:18429-18433.

913. Cytosolic long-chain acyl-CoA thioester hydrolase (Acyl-CoA_hydro)

This family consist of various cytosolic long-chain acyl-CoA thioester hydrolases including human and rat [1,2]. The aligned region is repeated with in the sequence of human and rat cytosolic long-chain acyl-CoA thioester hydrolases of this family. Long-chain acyl-CoA hydrolases hydrolyse palmitoyl-CoA to CoA and palmitate, they also catalyse the hydrolysis of other long chain fatty acyl-CoA thioesters. Long-chain acyl-CoA hydrolases are present in all living organisms and they may provide a mechanism for the control of lipid metabolism [1].

Number of members: 24

[1] Yamada J, Furihata T, Iida N, Watanabe T, Hosokawa M, Satoh T, Someya A, Nagaoka I, Suga T; Medline: 97236308 "Molecular cloning and expression of cDNAs encoding rat brain and liver cytosolic long-chain acyl-CoA hydrolases." Biochem Biophys Res Commun 1997;232:198-203.

[2] Broustas CG, Larkins LK, Uhler MD, Hajra AK; Medline: 96209964 "Molecular cloning and expression of cDNA encoding rat brain cytosolic acyl-coenzyme A thioester hydrolase." J Biol Chem 1996;271:10470-10476.

914. Agglutinin

Lectin (probable mannose binding)

Members of this family are plant lectins. Many if not all are mannose specific.

Number of members: 87

[1] Wright CS, Hester G; Medline: 97094989 "The 2.0 Å structure of a cross-linked complex between snowdrop lectin and a branched mannopentaose: evidence for two unique binding modes." Structure 1996;4:1339-1352.

5 915. (ANF_RECEPTORS)

Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.

10

Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction.

15

GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short.

25

A pattern was developed from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain.

30

Consensus pattern G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Garbers D.L. New Biol. 2:499-504(1990).

[2] Schulz S., Chinkers M., Garbers D.L. FASEB J. 2:2026-2035(1989).

916. (Apocytochrome)

Cytochrome c family heme-binding site signature

5 In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and
 10 reaction center cytochrome c.

Consensus pattern C-{CPWHF}{CPWHF SEQ ID NO:193}}-{CPWR}{CPWR SEQ ID NO:194}}-C-H-{CFYW}{CFYW SEQ ID NO:195}} Sequences known to belong to this class detected by the pattern ALL, except for four cytochrome c's which lack the first
 15 thioether bond. Other sequence(s) detected in SWISS-PROT454.

Note: some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !

20 [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

917. ATP-synt_A-c. ATP synthase Alpha chain, C terminal

[1] Medline: 94344236. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Abrahams JP, Leslie AG, Lutter R, Walker JE; Nature 1994;370:621-628.

25 Number of members: 125

918. (Basic)

Myc-type, 'helix-loop-helix' dimerization domain signature

HELIX_LOOP_HELIX

30

A number of eukaryotic proteins, which probably are sequence specific DNA- binding proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid residues. It has been proposed [1] that this domain is formed of two amphipathic helices

joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH) domain mediates protein dimerization and has been found in the proteins listed below [2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid residues that is adjacent to the HLH domain and specifically binds to DNA. They are referred

5 as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on the core sequence 'CANNTG', also referred to as the E-box motif. The homo- or heterodimerization mediated by the HLH domain is independent of, but necessary for DNA binding, as two basic regions are required for DNA binding activity. The HLH proteins

10 lacking the basic domain (Emc, Id) function as negative regulators since they form heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also repress transcription although they can bind DNA. The proteins of this subfamily act together with co-repressor proteins, like groucho, through their C-terminal motif WRPW.

- The myc family of cellular oncogenes [4], which is currently known to contain four

15 members: c-myc [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular differentiation and proliferation.

- Proteins involved in myogenesis (the induction of muscle cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or herculin), in birds CMD1 (QMF-1), in *Xenopus* MyoD and MF25, in *Caenorhabditis elegans* CeMyoD, and in *Drosophila*

20 *nautilus* (nau).

- Vertebrate proteins that bind specific DNA sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB.

- Vertebrate neurogenic differentiation factor 1 that acts as differentiation factor during

25 neurogenesis.

- Vertebrate MAX protein, a transcription regulator that forms a sequence- specific DNA-binding protein complex with myc or mad.

- Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

30 - Proteins of the bHLH/PAS superfamily which are transcriptional activators. In mammals, AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and

human BMAL1. In drosophila, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA).

- Mammalian transcription factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box.

5 - Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a symmetrical DNA sequence that is found in a variety of viral and cellular promoters.

- Human lyl-1 protein; which is involved, by chromosomal translocation, in T- cell leukemia.

10 - Human transcription factor AP-4.

- Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47.

- Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal
15 translocation, in stem-cell leukemia.

- Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a basic DNA-binding domain but are able to form heterodimers with other HLH proteins, thereby inhibiting binding to DNA.

- Drosophila extra-macrochaetae (emc) protein, which participates in sensory organ
20 patterning by antagonizing the neurogenic activity of the achaete- scute complex. Emc is the homolog of mammalian Id proteins.

- Human Sterol Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes.

25 - Drosophila achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system.

- Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins.

- Drosophila atonal protein (ato) which is involved in neurogenesis.

30 - Drosophila daughterless (da) protein, which is essential for neurogenesis and sex-determination.

- Drosophila deadpan (dpn), a hairy-like protein involved in the functional differentiation of neurons.

- *Drosophila hairy (h)* protein, a transcriptional repressor which regulates the embryonic segmentation and adult bristle patterning.

- *Drosophila* twist (twi) protein, which is involved in the establishment of germ layers in embryos.

10 - Yeast centromere-binding protein 1 (CPF1 or CBF1). This protein is involved in chromosomal segregation. It binds to a highly conserved DNA sequence, found in centromeres and in several promoters.

15 - Yeast phosphate system positive regulatory protein PHO4 which interacts with the upstream activating sequence of several acid phosphatase genes.

- *Neurospora crassa* nuc-1, a protein that activates the transcription of structural genes for phosphorus acquisition.

20

XXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----XXXXXXXXXXXXXXXXXXXXXXXXXXXXX Amphipathic
helix 1 Loop Amphipathic helix 2

25 The signature pattern that had been developed to detect this domain spans completely the second amphipathic helix.

30 [LIVMAGSNT][LIVMAGSNT SEQ ID NO:307]]-{FYWCPHKR}_{FYWCPHKR SEQ ID
NO:308))-[LIVMT][LIVMT SEQ ID NO:1]]-[LIVM][LIVM SEQ ID NO:4]]-x(2)-
[STAV][STAV SEQ ID NO:105]]-[LIVMSTACKR][LIVMSTACKR SEQ ID NO:309]]-x-
[VMFYH][VMFYH SEQ ID NO:310]]-[LIVMTA][LIVMTA SEQ ID NO:311]]-{P}-{P}-
[LIVMRKHQ][LIVMRKHQ SEQ ID NO:312]] Sequences known to belong to this class

detected by the pattern the majority but far from all. Other sequence(s) detected in SWISS-PROT135.

[1] Murre C., McCaw P.S., Baltimore D. Cell 56:777-783(1989).

5 [2] Garrel J., Campuzano S. BioEssays 13:493-498(1991).

[3] Kato G.J., Dang C.V. FASEB J. 6:3065-3072(1992).

[4] Krause M., Fire A., Harrison S.W., Priess J., Weintraub H. Cell 63:907-919(1990).

[5] Riechmann V., van Cruechten I., Sablitzky F. Nucleic Acids Res. 22:749-755(1994).

10 919. (Beta-lactamase)

Beta-lactamases classes -A, -C, and -D active site

Beta-lactamases (EC 3.5.2.6) [1,2] are enzymes which catalyze the hydrolysis of an amide bond in the beta-lactam ring of antibiotics belonging to the penicillin/cephalosporin family. Four kinds of beta-lactamase have been identified [3]. Class-B enzymes are zinc containing proteins whilst class -A, C and D enzymes are serine hydrolases. The three classes of serine beta-lactamases are evolutionary related and belong to a superfamily [4] that also includes DD-peptidases and a variety of other penicillin-binding proteins (PBP's). All these proteins contain a Ser-x-x-Lys motif, where the serine is the active site residue. Although clearly homologous, the sequences of the three classes of serine beta-lactamases exhibit a large degree of variability and only a small number of residues are conserved in addition to the catalytic serine.

25 Since a pattern detecting all serine beta-lactamases would also pick up many unrelated sequences, it was decided to provide specific patterns, centered on the active site serine, for each of the three classes.

30 Consensus pattern [FY]-x-[LIVMFY][LIVMFY SEQ ID NO:18])-x-S-[TV]-x-K-x(4)-[AGLM][AGLM SEQ ID NO:739])-x(2)-[LC] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-A beta-lactamases. Other sequence(s) detected in SWISS-PROT7.

770

Consensus pattern F-E-[LIVM][LIVM SEQ ID NO:4]-G-S-[LIVMG][LIVMG SEQ ID NO:202]-[SA]-K [The first S is the active site residue] Sequences known to belong to this class detected by the patternALL class-C beta-lactamases. Other sequence(s) detected in SWISS-PROTNONE.

5

Consensus pattern [PA]-x-S-[ST]-F-K-[LIV]-[PAL]-x-[STA]-[LI] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-D beta-lactamases. Other sequence(s) detected in SWISS-PROTNONE.

- 10 [1] Ambler R.P. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 289:321-331(1980).
 [2] Pastor N., Pinero D., Valdes A.M., Soberon X. Mol. Microbiol. 4:1957-1965(1990).
 [3] Bush K. Antimicrob. Agents Chemother. 33:259-263(1989).
 [4] Joris B., Ghuysen J.-M., Dive G., Renard A., Dideberg O., Charlier P., Frere J.M., Kelly J.A., Boyington J.C., Moews P.C., Knox J.R. Biochem. J. 250:313-324(1988).

15

920. Biotin protein ligase (BPL)

Biotin is covalently attached at the active site of certain enzymes that transfer carbon dioxide from bicarbonate to organic acids to form cellular metabolites. Biotin protein ligase (BPL) is the enzyme responsible for attaching biotin to a specific lysine at the active site of biotin enzymes. Each organism probably has only one BPL. Biotin attachment is a two step reaction that results in the formation of an amide linkage between the carboxyl group of biotin and the epsilon-amino group of the modified lysine [2].

Number of members: 26

25

- [1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Medline: 93028443 "Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains." Proc Natl Acad Sci USA 1992;89:9257-9261.
 [2] Chapman-Smith A, Cronan JE Jr; Medline: 10470036 "The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity." Trends Biochem Sci 1999;24:359-363.

30

921. (BRCA2_repeat)

The alignment covers only the most conserved region of the repeat. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature

- 5 [1] Bork P, Blomberg N, Nilges M; Medline: 96241568 "Internal repeats in the BRCA2 protein sequence." Nat Genet 1996;13:22-23.

Number of members: 63

10 922. (C6)

This domain of unknown function is found in the *C. elegans* protein Swiss:Q19522. It is presumed to be an extracellular domain. The C6 domain contains six conserved cysteine residues in most copies of the domain. However some copies of the domain are missing
15 cysteine residues 1 and 3 suggesting that these form a disulphide bridge.

Number of members: 23

923. Cadherin cytoplasmic region (Cadherin_C_term)

- 20 Cadherins are vital in cell-cell adhesion during tissue differentiation. Cadherins are linked to the cytoskeleton by catenins. Catenins bind to the cytoplasmic tail of the cadherin. Cadherins cluster to form foci of homophilic binding units. A key determinant to the strength of the binding that it is mediated by cadherins is the juxtamembrane region of the cadherin. This region induces clustering and also binds to the protein p120ctn [1].

25 Number of members: 59

[1] Yap AS, Niessen CM, Gumbiner BM; Medline: 98234411 "The juxtamembrane region of the cadherin cytoplasmic tail supports lateral clustering, adhesive strengthening, and interaction with p120ctn." J Cell Biol 1998;141:779-789.

- 30 [2] Barth AI, Nathke IS, Nelson WJ; Medline: 97471931 "Cadherins, catenins and APC protein: interplay between cytoskeletal complexes and signaling pathways." Curr Opin Cell Biol 1997;9:683-690.

[3] Braga VM, Machesky LM, Hall A, Hotchin NA; Medline: 97327766 "The small GTPases Rho and Rac are required for the establishment of cadherin-dependent cell-cell contacts." J Cell Biol 1997;137:1421-1431.

5 924. Clathrin propeller repeat (Clathrin_propel)

Clathrin is the scaffold protein of the basket-like coat that surrounds coated vesicles. The soluble assembly unit, a triskelion, contains three heavy chains and three light chains in an extended three-legged structure. Each leg contains one heavy and one light chain. The N-terminus of the heavy chain is known as the globular domain, and is composed of seven repeats which form a beta propeller [1].

Number of members: 61

[1] ter Haar E, Musacchio A, Harrison SC, Kirchhausen T; Medline: 99043510 "Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker." Cell. 1998;95:563-573.

925. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature (complex1_30Kd)

20 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 30 Kd (in mammals) which has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.
- Mitochondrial encoded in *Paramecium* (protein P1), and in the slime mold *Dictyostelium discoideum* (ORF 209).
- 30 - Chloroplast encoded in various higher plants (ORF 159). It is also present in bacteria:
 - In the cyanobacteria *Synechocystis* strain PCC 6803 (gene *ndhJ*).
 - Subunit C of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoC*).
 - Subunit NQO5 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.

This protein, in its mature form, consists of from 157 to 266 amino acid residues. The best conserved region is located in the C-terminal section and can be used as a signature pattern.

- 5 Consensus pattern E-R-E-x(2)-[DE]-[LIVMFY][LIVMFY SEQ ID NO:18]](2)-x(6)-[HK]-x(3)-[KRP]-x-[LIVM][LIVM SEQ ID NO:4]]-[LIVMYS][LIVMYS SEQ ID NO:740]] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

- 10 [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).
[2]Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

926. Respiratory-chain NADH dehydrogenase 49 Kd subunit signature (complex1_49Kd)

- 15 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 49 Kd (in mammals), which is the third largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 4Fe-4S iron-sulfur cluster. The 49 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.
25 - Mitochondrial encoded in protozoan such as *Paramecium* (ORF 400), *Leishmania* and *Trypanosoma* (MURF 3).
- Chloroplast encoded in various higher plants (ORF 392).

The 49 Kd subunit is highly similar to [3,4]:

- Subunit D of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoD*).
30 - Subunit NQO4 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- Subunit 5 of *Escherichia coli* formate hydrogenlyase (gene *hycE*).
- Subunit G of *Escherichia coli* hydrogenase-4 (gene *hyfG*).

A highly conserved region was selected as signature pattern, located in the N-terminal section of this subunit.

Consensus pattern [LIVMH][LIVMH SEQ ID NO:703]-H-[RT]-[GA]-x-E-K-

5 [LIVMTN][LIVMTN SEQ ID NO:280]-x-E-x-[KRQ] Sequences known to belong to this class detected by the patternALL.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

10 [3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

927. (COX2)

15

Cytochrome c oxidase (EC 1.9.3.1) [1,2] is an oligomeric enzymatic complex which is a component of the respiratory chain and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. The enzyme

20 complex consists of 3-4 subunits (prokaryotes) to up to 13 polypeptides (mammals).

Subunit 2 (CO II) transfers the electrons from cytochrome c to the catalytic subunit 1. It contains two adjacent transmembrane regions in its N-terminus and the major part of the protein is exposed to the periplasmic or to the mitochondrial intermembrane space, respectively. CO II provides the substrate-binding site and contains a copper center called Cu(A), probably the primary acceptor in cytochrome c oxidase. An exception is the

25 corresponding subunit of the cbb3-type oxidase which lacks the copper A redox-center. Several bacterial CO II have a C-terminal extension that contains a covalently bound heme c.

30 It has been shown [3,4] that nitrous oxide reductase (EC 1.7.99.6) (gene nosZ) of Pseudomonas has sequence similarity in its C-terminus to CO II. This enzyme is part of the bacterial respiratory system which is activated under anaerobic conditions in the presence of nitrate or nitrous oxide. NosZ is a periplasmic homodimer that contains a dinuclear copper

center, probably located in a 3- dimensional fold similar to the cupredoxin-like fold that has been suggested for the copper-binding site of CO II [3].

The dinuclear purple copper center is formed by 2 histidines and 2 cysteines [5]. This region was used as a signature pattern. The conserved valine and the conserved methionine are said to be involved in stabilizing the copper-binding fold by interacting with each other.

Consensus pattern V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M [The two C's and two H's are copper ligands] Sequences known to belong to this class detected by the patternALL, except for *Paramecium primaurelia* as well as in some plants where the pattern ends with Thr; an RNA editing event at this position could change this Thr to Met.

Note: cytochrome cbb(3) subunit 2 does not belong to this family.

[1] Capaldi R.A., Malatesta F., Darley-USmar V.M. *Biochim. Biophys. Acta* 726:135-148(1983).

[2] Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B. *J. Bacteriol.* 176:5587-5600(1994).

[3] van der Oost J., Lappalainen P., Musacchio A., Warne A., Lemieux L., Rumbley J., Gennis R.B., Aasa R., Pascher T., Malmstrom B.G., Saraste M. *EMBO J.* 11:3209-3217(1992).

[4] Zumft W.G., Dreutsch A., Loechele S., Cuypers H., Friedrich B., Schneider B. *Eur. J. Biochem.* 208:31-40(1992).

928. Cytochrome C assembly protein (CytC_asm)

This family consists of various proteins involved in cytochrome c assembly from mitochondria and bacteria; CycK from *Rhizobium* [3], CcmC from *E. coli* and *Paracoccus denitrificans* [2,1] and orf240 from wheat mitochondria [4]. The members of this family are probably integral membrane proteins with six predicted transmembrane helices. It has been proposed that members of this family comprise a membrane component of an ABC (ATP binding cassette) transporter complex. It is also proposed that this transporter is necessary for transport of some component needed for cytochrome c assembly. One member CycK

contains a putative heme-binding motif [3], orf240 also contains a putative heme-binding motif and is a proposed ABC transporter with c-type heme as its proposed substrate [4]. However it seems unlikely that all members of this family transport heme nor c-type apocytochromes because CcmC in the putative CcmABC transporter transports neither [1].

5 Number of members: 67

[1] Page D, Pearce DA, Norris HA, Ferguson SJ; Medline: 97195802 "The *Paracoccus denitrificans* ccmA, B and C genes: cloning and sequencing, and analysis of the potential of their products to form a haem or apo-c-type cytochrome transporter. MICROBIOLOGY
10 1997;143:563-576.

[2] Thoeny-meyer L, Fischer F, Kunzler P, Ritz D, Hennecke H; Medline: 95362656 "Escherichia coli genes required for cytochrome c maturation." J. BACTERIOL
1995;177:4321-4326.

[3] Delgado MJ, Yeoman KH, Wu G, Vargas C, Davies A, Poole RK, Johnston AWB,
15 Downie JA; Medline: 95394794 "Characterization of the cychJKL genes involved in cytochrome c biogenesis and symbiotic nitrogen fixation in *Rhizobium leguminosarum*." J. BACTERIOL 1995;177:4927-4934.

[4] Bonnard G, Grienemberger JM; Medline: 95124303 "A gene proposed to encode a transmembrane domain of an ABC transporter is expressed in wheat mitochondria." MOL.
20 GEN. GENET 1995;246:91-99.

929. Cytochrome b559 subunits heme-binding site signature (cytochr_b559)

Cytochrome b559 [1] is an essential component of photosystem II complex from oxygenic
25 photosynthetic organisms. It is an integral thylakoid membrane protein composed of two subunits, alpha (gene psbE) and beta (gene psbF), each of which contains a histidine residue located in a transmembrane region. The two histidines coordinate the heme iron of cytochrome b559.

30 The region around the heme-binding residue of both subunits is very similar and can be used as a signature pattern.

777

Consensus pattern[LIV]-x-[ST]-[LIVF][LIVE SEQ ID NO:127]]-R-[FYW]-x(2)-[IV]-H-
[STGA][STGA SEQ ID NO:741]]-[LIV]-[STGA][STGA SEQ ID NO:741]]-[IV]-P [H is the
heme iron ligand] Sequences known to belong to this class detected by the patternALL. Other
sequence(s) detected in SWISS-PROT NONE.

5

[1] Pakrasi H.B., de Ciechi P., Whitmarsh J. EMBO J. 10:1619-1627(1991).

930. Cytochrome b/b6 signatures (Cytochrome_b)

10

In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component
of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-
cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein,
cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1),
also known as the b6f complex.

15

Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid
residues that probably has 8 transmembrane segments. In plants and cyanobacteria,
cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence
of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD
corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups,
known as b562 and b566. Four conserved histidine residues are postulated to be the ligands
of the iron atoms of these two heme groups.

20

Apart from regions around some of the histidine heme ligands, there are a few conserved
regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-
E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments.
It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo
(where o stands for outside) - located on the outer side of the membrane.

25

30

A schematic representation of the structure of cytochrome b/b6 is shown below.

+---Fe-b562----+ | +---Fe-b566--|-+ ||||

xxxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxxxPEWxxxxxxxxxxxxxxxxxxxxx <-----
 ---Cytochrome-b-----> <----Cytochrome-b6-petB-----><--Cytochrome-
 b6-petD----->

5

Two signature patterns were developed for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.

10 Consensus pattern [DENQ][DENQ SEQ ID NO:371]-x(3)-G-[FYWMQ][FYWMQ SEQ ID NO:742]-x-[LIVMF][LIVMF SEQ ID NO:2]-R-x(2)-H [H is a heme b562 ligand]
 Sequences known to belong to this class detected by the patternALL, except for 5 sequences.

15 Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the patternALL, except for *Odocoileus hemionus* (mule deer) and *Paramecium tetraurelia* cytochrome b.

[1] Howell N. J. Mol. Evol. 29:157-169(1989).

20 [2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim. Biophys. Acta 1143:243-271(1993).

931. Phorbol esters / diacylglycerol binding domain (DAG_PE-bind)

25 Diacylglycerol (DAG) is an important second messenger. Phorbol esters (PE) are analogues of DAG and potent tumor promoters that cause a variety of physiological changes when administered to both cells and tissues. DAG activates a family of serine/threonine protein kinases, collectively known as protein kinase C (PKC) [1]. Phorbol esters can directly stimulate PKC. The N- terminal region of PKC, known as C1, has been shown [2] to bind PE and DAG in a phospholipid and zinc-dependent fashion. The C1 region contains one or two
 30 copies (depending on the isozyme of PKC) of a cysteine-rich domain about 50 amino-acid residues long and essential for DAG/PE-binding. Such a domain has also been found in the following proteins:

- Diacylglycerol kinase (EC 2.7.1.107) (DGK) [3], the enzyme that converts DAG into phosphatidate. It contains two copies of the DAG/PE-binding domain in its N-terminal section. At least five different forms of DGK are known in mammals.

- N-chimaerin. A brain specific protein which shows sequence similarities with the BCR protein at its C-terminal part and contains a single copy of the DAG/PE-binding domain at its N-terminal part. It has been shown [4,5] to be able to bind phorbol esters.

- The raf/mil family of serine/threonine protein kinases. These protein kinases contain a single N-terminal copy of the DAG/PE-binding domain.

- The unc-13 protein from *Caenorhabditis elegans*. Its function is not known but it contains a copy of the DAG/PE-binding domain in its central section and has been shown to bind specifically to a phorbol ester in the presence of calcium [6].

- The vav oncogene. Vav was generated by a genetic rearrangement during gene transfer assays. Its expression seems to be restricted to cells of hematopoietic origin. Vav seems to contain a DAG/PE-binding domain in the central part of the protein.

- The *Drosophila* GTPase activating protein rotund.

The DAG/PE-binding domain binds two zinc ions; the ligands of these metal ions are probably the six cysteines and two histidines that are conserved in this domain. A signature pattern was developed that spans completely the DAG/PE domain.

Consensus pattern H-x-[LIVMFYW][LIVMFYW SEQ ID NO:26]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC][LIVMFC SEQ ID NO:90]-x(5,10)-C-x(2)-C-x(4)-[HD]-x(2)-C-x(5,9)-C [All the C and H are involved in binding Zinc] Sequences known to belong to this class detected by the pattern ALL, except a few DGK's.

[1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

[2] Ono Y., Fujii T., Igarashi K., Kuno T., Tanaka C, Kikkawa U., Nishizuka Y. Proc. Natl. Acad. Sci. U.S.A. 86:4868-4871(1989).

[3] Sakane F., Yamada K., Kanoh H., Yokoyama C., Tanabe T. Nature 344:345-348(1990).

[4] Ahmed S., Kozma R., Monfries C., Hall C., Lim H.H., Smith P., Lim L. Biochem. J. 272:767-773(1990).

[5] Ahmed S., Kozma R., Lee J., Monfries C., Harden N., Lim L. Biochem. J. 280:233-241(1991).

[6] Ahmed S., Maruyama I.N., Kozma R., Lee J., Brenner S., Lim L. Biochem. J. 287:995-999(1992).

[7] Boguski M.S., Bairoch A., Attwood T.K., Michaels G.S. Nature 358:113-113(1992).

5 932. 3-dehydroquinate synthase (DHQ_synthase)

[1] Barten R, Meyer TF; Medline: 98273626 "Cloning and characterisation of the Neisseria gonorrhoeae aroB gene." Mol Gen Genet 1998;258:34-44.

10 [2] Hawkins AR, Lamb HK; Medline: 96048023 "The molecular biology of multidomain proteins. Selected examples." Eur J Biochem 1995;232:7-18.

The 3-dehydroquinate synthase EC:4.6.1.3 domain is present in isolation in various bacterial 3-dehydroquinate synthases and also present as a domain in the pentafunctional AROM polypeptide Swiss:P07547 [2]. 3-dehydroquinate (DHQ) synthase catalyses the formation of
15 dehydroquinate (DHQ) and orthophosphate from 3-deoxy-D-arabino heptulosonic 7 phosphate [1]. This reaction is part of the shikimate pathway which is involved in the biosynthesis of aromatic amino acids.

Number of members: 25

20 933. Dihydrofolate reductase signature (DiHfolate_red)

Dihydrofolate reductases (EC 1.5.1.3) [1] are ubiquitous enzymes which catalyze the reduction of folic acid into tetrahydrofolic acid. They can be inhibited by a number of antagonists such as trimethoprim and methotrexate which are used as antibacterial or
25 anticancerous agents. A signature pattern was derived from a region in the N-terminal part of these enzymes, which includes a conserved Pro-Trp dipeptide; the tryptophan has been shown [2] to be involved in the binding of substrate by the enzyme.

30 Consensus pattern[LVAGC][LVAGC SEQ ID NO:743)]-[LIF]-G-x(4)-[LIVMF][LIVMF
SEQ ID NO:2)]-P-W-x(4,5)-[DE]-x(3)-[FYIV][FYIV SEQ ID NO:744)]-
x(3)-[STIQ][STIQ SEQ ID NO:745)] Sequences known to belong to this class detected by

the patternALL, except for type II bacterial, plasmid-encoded, dihydrofolate reductases which do not belong to the same class of enzymes.

[1] Harpers' Review of Biochemistry, Lange, Los Altos (1985).

[2] Bolin J.T., Filman D.J., Matthews D.A., Hamlin R.C., Kraut J. J. Biol. Chem. 257:13650-13662(1982).

5

934. (DIL)

[1] Ponting CP; Medline: 95397417 "AF-6/cno: neither a kinesin nor a myosin, but a bit of both." Trends Biochem Sci 1995;20:265-266.

10

Number of members: 31

935. (DNA_gyraseB_C)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

15

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

20

Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE).

25

In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

30

<-----About-1400-residues----->

[-----Protein 39-*-----][----Protein 52----] Phage T4

[-----gyrB-----*-----][-----gyrA-----] Prokaryote II

Archaeobacteria

[-----parE-----*-----][-----parD-----] Prokaryote IV

[-----*-----] Eukaryote and ASF

'*': Position of the pattern.

5

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

10 Consensus pattern [LIVMA][LIVMA SEQ ID NO:30]-x-E-G-[DN]-S-A-x-[STAG][STAG SEQ ID NO:20] Sequences known to belong to this class detected by the pattern ALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

15 [3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

936. (DNA_topoisolIV)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

20

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

Topoisomerase II is found in phages, archaeobacteria, prokaryotes, eukaryotes, and in African

25 Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaeobacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE).

30 In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```

5  <-----About-1400-residues----->
    [-----Protein 39-*-----][----Protein 52----] Phage T4
    [-----gyrB-----*-----][-----gyrA-----] Prokaryote II Archaeobacteria
    [-----parE-----*-----][-----parD-----] Prokaryote IV
    [-----*-----] Eukaryote and ASF
10  '*': Position of the pattern.

```

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

```

15  Consensus pattern [LIVMA][LIVMA SEQ ID NO:30])-x-E-G-[DN]-S-A-x-[STAG][STAG
    SEQ ID NO:20)] Sequences known to belong to this class detected by the patternALL.

```

- [1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).
- 20 [2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).
- [3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).
- [4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

937. Prolyl oligopeptidase family serine active site (DPPIV_N_term)

```

25  The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related
    peptidases whose catalytic activity seems to be provided by a charge relay system similar to
    that of the trypsin family of serine proteases, but which evolved by independent convergent
    evolution. The known members of this family are listed below.

```

- 30 - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is
 an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence
 of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium

meningosepticum and *Aeromonas hydrophila*); there is a high degree of sequence conservation between these sequences.

- *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.

5 - Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

10 - Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

15 A conserved serine residue has experimentally been shown (in *E.coli* protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

20

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW][LIVMFYW SEQ ID NO:26]-x(14)-G-x-S-x-G-G-[LIVMFYW][LIVMFYW SEQ ID NO:26](2) [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A.

25 Note: these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).

30 [3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

938. Deoxyhypusine synthase (DS)

Eukaryotic initiation factor 5A (eIF-5A) contains an unusual amino acid, hypusine [N epsilon-(4-aminobutyl-2-hydroxy)lysine]. The first step in the post-translational formation of hypusine is catalysed by the enzyme deoxyhypusine synthase (DS) EC:1.1.1.249. The modified version of eIF-5A, and DS, are required for eukaryotic cell proliferation [1].

Number of members: 9

[1] Liao DI, Wolff EC, Park MH, Davies DR; Medline: 98154315 "Crystal structure of the NAD complex of human deoxyhypusine synthase: an enzyme with a ball-and-chain mechanism for blocking the active site." Structure 1998;6:23-32.

939. (DUF21)

Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

940. (DUF59)

This family includes prokaryotic proteins of unknown function. The family also includes PhaH Swiss:O84984 from *Pseudomonas putida*. PhaH forms a complex with PhaF Swiss:O84982, PhaG Swiss:O84983 and PhaI Swiss:O84985, which hydroxylates phenylacetic acid to 2-hydroxyphenylacetic acid [1]. So members of this family may all be components of ring hydroxylating complexes.

Number of members: 15

[1] Olivera ER, Minambres B, Garcia B, Muniz C, Moreno MA, Ferrandez A, Diaz E, Garcia JL, Luengo JM; Medline: 98263372 "Molecular characterization of the phenylacetic acid

catabolic pathway in *Pseudomonas putida* U: the phenylacetyl-CoA catabolon." *Proc Natl Acad Sci U S A* 1998;95:6419-6424.

941. (DUF82)

5

The protein contains four conserved cysteines that may be involved in metal binding or disulphide bridges.

Number of members: 4

10 942. Riboflavin kinase / FAD synthetase (FAD_Synth)

This family consists part of the bifunctional enzyme riboflavin kinase / FAD synthetase.

These enzymes have both ATP:riboflavin 5'-phospho transferase and ATP:FMN-

adenylyltransferase activities [1]. They catalyse the 5'-phosphorylation of riboflavin to FMN

15 and the adenylation of FMN to FAD [1].

CAUTION: It is not clear if this region of the enzymes catalyses either or both of the enzymatic reactions.

Number of members: 27

20 [1] Manstein DJ, Pai EF; Medline: 87057286 "Purification and characterization of FAD synthetase from *Brevibacterium ammoniagenes*." *J Biol Chem* 1986;261:16169-16173.

943. [2Fe-2S] binding domain (fer2_2)

25 [1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*." *Science* 1995;270:1170-1176.

Number of members: 53

30 944. Filovirus glycoprotein (Filo_glycop)

This family includes an extracellular region from the envelope glycoprotein of Ebola and Marburg viruses. This region is also produced as a separate transcript that gives rise to a non-

structural, secreted glycoprotein, which is produced in large amounts and has an unknown function [1]. Processing of this protein may be involved in viral pathogenicity [2].

Number of members: 23

- 5 [1] Volchkov VE, Feldmann H, Volchkova VA, Klenk HD; Medline: 98245155 "Processing of the Ebola virus glycoprotein by the proprotein convertase furin." Proc Natl Acad Sci U S A 1998;95:5762-5767.

- [2] Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST; Medline: 96195018 "The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed
10 through transcriptional editing." Proc Natl Acad Sci U S A 1996;93:3602-3607.

945. Frataxin-like domain (Frataxin_Cyay)

This family contains proteins that have a domain related to the globular C-terminus of
15 Frataxin the protein that is mutated in Friedreich's ataxia. This domain is found in a family of bacterial proteins. The function of this domain is currently unknown.

Number of members: 12

- [1] Gibson TJ, Koonin EV, Musco G, Pastore A, Bork P; Medline: 97084946 "Friedreich's
20 ataxia protein: phylogenetic evidence for mitochondrial dysfunction." Trends Neurosci 1996;19:465-468.

946. (GAF)

25 Domain present in phytochromes and cGMP-specific phosphodiesterases.

Number of members: 296

- [1] Aravind L, Ponting CP; Medline: 98094688 "The GAF domain: an evolutionary link
between diverse phototransducing proteins." Trends Biochem Sci 1997;22:458-459.

30

947. Galaptin signature (Gal-bind_lectin)

All vertebrates synthesize soluble galactoside-binding lectins [1,2,3] (also known as galectins, galaptins or S-lectin). These carbohydrate-binding proteins are developmentally regulated. Although their exact physiological role is not yet clear they seem to be involved in differentiation, cellular regulation and tissue construction. The sequence of galactoside-binding lectins from electric eel (electrolectin), conger eel (congerin), chicken and a number of mammalian species is known. These lectins are proteins of about 130 to 140 amino acid residues (14 Kd to 16 Kd).

A number of other proteins are known to belong to this family:

- Galectin-3 (also known as MAC-2 antigen; CBP-35 or IgE-binding protein), a 35 Kd lectin which binds immunoglobulin E and which is composed of two domains: a N-terminal domain that consist of tandem repeats of a glycine/ proline-rich sequence and a C-terminal galaptin domain.

- Galectin-4 [4], which is composed of two galaptin domains.

- Galectin-5.

- Galectin-7 [5], a keratinocyte protein which could be involved in cell-cell and/or cell-matrix interactions necessary for normal growth control.

- Galectin-8 [6], which is composed of two galaptin domains.

- Galectin-9 [7], which is composed of two galaptin domains.

- Human eosinophil lysophospholipase (EC 3.1.1.5) [8] (Charcot-Leyden crystal protein), a protein that may have both an enzymatic and a lectin activities. It forms hexagonal bipyramidal crystals in tissues and secretions from sites of eosinophil-associated inflammation.

- *Caenorhabditis elegans* 32 Kd lactose-binding lectin [9]. This lectin is composed of two galaptin domains.

- *Caenorhabditis elegans* lec-7 and lec-8.

One of the conserved regions of these lectins contains a tryptophan that has been shown [10] to be essential to the binding of galactosides. This region was used as a signature pattern for these proteins.

Consensus pattern W-[GEK]-x-[EQ]-x-[KRE]-x(3,6)-[PCTF][PCTF SEQ ID NO:746]-[LIVMF][LIVMF SEQ ID NO:2]-[NQEGSKV][NQEGSKV SEQ ID NO:747]-x-[GH]-x(3)-[DENKHS][DENKHS SEQ ID NO:748]-[LIVMFC][LIVMFC SEQ ID NO:90] [W

binds carbohydrate] Sequences known to belong to this class detected by the pattern ALL, except for pig galectin 4.

- [1] Barondes S.H., Gitt M.A., Leffler H., Cooper D.N.W. *Biochimie* 70:1627-1632(1988).
- 5 [2] Hirabayashi J., Kasai K.-I. *J. Biochem.* 104:1-4(1988).
- [3] Barondes S.H., Castronovo V., Cooper D.N.W., Cummings R.D., Drickamer K., Feizi T., Gitt M.A., Hirabayashi J., Hughes C., Kasai K.-I., Leffler H., Liu F.-T., Lotan R., Mercurio A.M., Monsigny M., Pillair S., Poirer F., Raz A., Rigby P.W.J., Rini J.M., Wang J.L. *Cell* 76:597-598(1994).
- 10 [4] Oda Y., Herrmann J., Gitt M., Turck C.W., Burlingame A.L., Barondes S.H., Leffler H. *J. Biol. Chem.* 268:5929-5939(1993).
- [5] Madsen P., Rasmussen H.H., Flint T., Gromov P., Kruse T.A., Honore B., Vorum H., Celis J.E. *J. Biol. Chem.* 270:5823-5829(1995).
- [6] Hadari Y.R., Paz K., Dekel R., Mestrovic T., Accili D., Zick Y. *J. Biol. Chem.* 270:3447-15 3453(1995).
- [7] Wada J., Kanwar Y.S. *J. Biol. Chem.* 272:6078-6086(1997).
- [8] Ackerman S.J., Corrette S.E., Rosenberg H.F., Bennett J.C., Mastrianni D.M., Nicholson-Weller A., Weller P.F., Chin D.T., Tenen D.G. *J. Immunol.* 150:456-468(1993).
- [9] Hirabayashi J., Satoh M., Kasai K.-I. *J. Biol. Chem.* 267:15485-15490(1992).
- 20 [10] Abbott W.M., Feizi T. *J. Biol. Chem.* 266:5552-5557(1991).

948. (GARS) Phosphoribosylglycinamide synthetase signature (phosphoribosylamine glycine ligase)

PROSITE: PDOC00164; cross-reference(s): PS00184

25 [1] catalyzes the second step in the de novo biosynthesis of purine, the ATP-dependent addition of 5-phosphoribosylamine to glycine to form 5'phosphoribosylglycinamide.

In bacteria GARS is a monofunctional enzyme (encoded by the *purD* gene), in of a bifunctional enzyme (encoded by the *ADE5,7* gene), in higher eukaryotes it is part, with 30 AIRS and with phosphoribosylglycinamide formyltransferase (GART) of a trifunctional enzyme (GARS-AIRS-GART).

The sequence of GARS is well conserved. A highly conserved octapeptide was selected as a signature pattern.

Consensus pattern R-F-G-D-P-E-x-[QM]

Sequences known to belong to this class detected by the pattern ALL.

- 5 [1] Aiba A., Mizobuchi K. J. Biol. Chem. 264:21239-21246(1989).

949. GLTT - GLTT repeat (12 copies)

This short repeat of unknown function is found in multiple copies in several *C. elegans* proteins. The repeat is five residues long and consists of XGLTT where X can be any amino
10 acid. Number of members: 34.

950. Glu_synthase - Conserved region in glutamate synthase

This family represents a region of the glutamate synthase protein. This region is expressed as a separate subunit in the glutamate synthase alpha subunit from archaebacteria, or part of a
15 large multidomain enzyme in other organisms. The aligned region of these proteins contains a putative FMN binding site and Fe-S cluster. Number of members: 44.

- [1] Medline: 97082505. Sequence of the GLT1 gene from *Saccharomyces cerevisiae* reveals the domain structure of yeast glutamate synthase. Filetici P, Martegani MP, Valenzuela L,
20 Gonzalez A, Ballario P; Yeast 1996;12:1359-1366.

951. (Glyco_hydro_2) Glycosyl hydrolases family 2 signatures

GLYCOSYL_HYDROL_F2_1; PS00608; GLYCOSYL_HYDROL_F2_2

It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of
25 sequence similarities, classified into a single family:

-Beta-galactosidases (EC 3.2.1.23) from bacteria such as *Escherichia coli* (genes *lacZ* and *ebgA*), *Clostridium acetobutylicum*, *Clostridium thermosulfurogenes*, *Klebsiella pneumoniae*, *Lactobacillus delbrueckii*, or *Streptococcus thermophilus* and from the fungi *Kluyveromyces lactis*.

30 -Beta-glucuronidase (EC 3.2.1.31) from *Escherichia coli* (gene *uidA*) and from mammals. One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [3], in *Escherichia coli lacZ*, to be the general acid/base catalyst in the active site of the enzyme. This region has been used as a signature pattern. A

highly conserved region located some sixty residues upstream from the active site glutamate has been selected as a second signature pattern.

Consensus pattern N-x-[LIVMFYWD][LIVMFYWD SEQ ID NO:299)]-R-

5 [STACN][STACN SEQ ID NO:300)](2)-H-Y-P-x(4)-[LIVMFYWS][LIVMFYWS SEQ ID NO:301)](2)-x(3)-[DN]-x(2)-G-[LIVMFYW][LIVMFYW SEQ ID NO:26)](4) Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [DENQLF][DENQLF SEQ ID NO:302)]-[KRVW][KRVW SEQ ID

10 NO:303)]-N-[HRY]-[STAPPV][STAPPV SEQ ID NO:749)]-[SAC]-[LIVMFS][LIVMFS SEQ ID NO:132)](3)-W-[GS]-x(2,3)-N-E [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for *Rhizobium meliloti lacZ*.

[1]Henrissat B. Biochem. J. 280:309-316(1991).

15 [2]Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. Gen. Microbiol. 137:369-380(1991).

[3]Gebler J.C., Aebersold R., Withers S.G. J. Biol. Chem. 267:11126-11130(1992).

952. (Glyco_hydro_3) Glycosyl hydrolases family 3 active site

20 PROSITE: PDOC00621. PROSITE cross-reference(s)PS00775; GLYCOSYL_HYDROL_F3

It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3), *Hansenula anomala*, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*, (BGL1 and BGL2),

25 *Schizophyllum commune* and *Trichoderma reesei* (BGL1).

-Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1), *Butyrivibrio fibrisolvens* (bglA), *Clostridium thermocellum* (bglB), *Escherichia coli* (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus albus*.

-*Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

30 -*Bacillus subtilis* hypothetical protein yzbA.

-*Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding *Haemophilus influenzae* protein.

One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta-glucosidase A3, to be implicated in the catalytic mechanism. This region was used as a signature pattern.

5 Consensus pattern[LIVM][LIVM SEQ ID NO:4](2)-[KR]-x-[EQK]-x(4)-G-
[LIVMFT][LIVMFT SEQ ID NO:282]-[LIVT][LIVT SEQ ID NO:165]-[LIVMF][LIVMF
[SEQ ID NO:2]-[ST]-D-x(2)-[SGADNH][SGADNI SEQ ID NO:283] [D is the active site
residue]

Sequences known to belong to this class detected by the patternALL.

10

[1]Henrissat B. Biochem. J. 280:309-316(1991).

[2]Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3]Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

15 953. GP120 - Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with Swiss:P01730 and a chemokine receptor on the cell surface. Number of members: 17891

20

[1]Medline: 98303379. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA; Nature 1998;393:648-659.

954. (GSP_{II}_E) Bacterial type II secretion system protein E signature

PROSITE: PDOC00567. PROSITE cross-reference(s) PS00662; T2SP_E

25

A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1,2], have been found to be evolutionary related. These proteins are listed below:

-The 'E' protein from the GSP operon of: *Aeromonas* (gene *exeE*); *Erwinia* (gene *outE*); *Escherichia coli* (gene *yheG*); *Klebsiella pneumoniae* (gene *pulE*); *Pseudomonas aeruginosa* (gene *xcpR*); *Vibrio cholerae* (gene *epsE*) and *Xanthomonas campestris* (gene *xpsE*).

30

-*Agrobacterium tumefaciens* Ti plasmid *virB* operon protein 11. This protein is required for the transfer of T-DNA to plants.

-*Bacillus subtilis* comG operon protein 1 which is required for the uptake of DNA by competent *Bacillus subtilis* cells.

-*Aeromonas hydrophila* tapB, involved in type IV pilus assembly.

-*Pseudomonas* protein pilB, which is essential for the formation of the pili.

5 -*Pseudomonas aeruginosa* protein twitching mobility protein pilT.

-*Neisseria gonorrhoeae* type IV pilus assembly protein pilF.

-*Vibrio cholerae* protein tcpT, which is involved in the biosynthesis of the tcp pilus.

-*Escherichia coli* protein hofB (hopB).

10 -*Escherichia coli* hypothetical protein ygcB.

-*Escherichia coli* hypothetical protein yggR.

These proteins have from 344 (pilT and virB11) to 568 (tapB) amino acids, they are probably cytoplasmically located and, on the basis of the presence of a conserved P-loop region (see <PDOC00017>), probably bind ATP. A region that overlaps the 'B' motif of

15 ATP-binding proteins was selected as a signature pattern.

Consensus pattern[LIVM][LIVM SEQ ID NO:4]-R-x(2)-P-D-x-[LIVM][LIVM SEQ ID NO:4]](3)-G-E-[LIVM][LIVM SEQ ID NO:4]-R-D

Sequences known to belong to this class detected by the patternALL, except for ygcB.

20

[1]Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2]Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

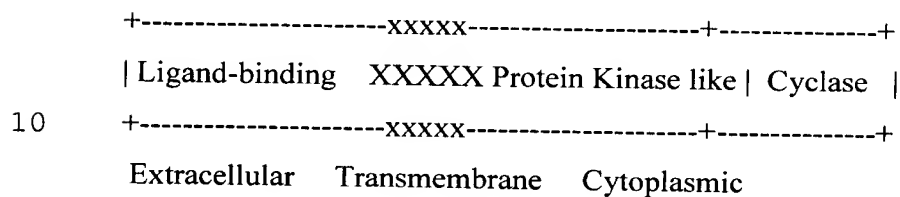
955. (guanylate_cyc) Guanylate cyclases signature

25 PROSITE: PDOC00425. PROSITE cross-reference(s) PS00452;

GUANYLATE_CYCLASES Guanylate cyclases (EC 4.6.1.2) [1 to 4] catalyze the formation of cyclic GMP (cGMP) from GTP. cGMP acts as an intracellular messenger, activating cGMP dependent kinases and regulating CGMP-sensitive ion channels. The role of cGMP as a second messenger in vascular smooth muscle relaxation and retinal photo-

30 transduction is well established. Guanylate cyclase is found both in the soluble and particular fraction of eukaryotic cells. The soluble and plasma membrane-bound forms differ in structure, regulation and other properties.

Most currently known plasma membrane-bound forms are receptors for small polypeptides. The topology of such proteins is the following: they have a N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain, followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain that appears important for proper signalling and a cyclase catalytic domain. This topology is schematically represented below.



The known guanylate cyclase receptors are:

- 15 -The sea-urchins receptors for speract and resact, which are small peptides that stimulate sperm motility and metabolism.
- The receptors for natriuretic peptides (ANF). Two forms of ANF receptors with guanylate cyclase activity are currently known: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP.
- 20 -The receptor for Escherichia coli heat-stable enterotoxin (GC-C). The endogenous ligand for this intestinal receptor seems to be a small peptide called guanylin.
- Retinal guanylate cyclase (retGC) which probably plays a specific functional role in the rods and/or cones of photoreceptors. It is not known if this protein acts as receptor, but its structure is similar to that of the other plasma membrane-bound GCs.

25 The soluble forms of guanylate cyclase are cytoplasmic heterodimers. The two subunits, alpha and beta are proteins of from 70 to 82 Kd which are highly related. Two forms of beta subunits are currently known: beta-1 which seems to be expressed in lung and brain, and beta-2 which is more abundant in kidney and liver.

30 The membrane and cytoplasmic forms of guanylate cyclase share a conserved domain which is probably important for the catalytic activity of the enzyme. Such a domain is also found twice in the different forms of membrane-bound adenylate cyclases (also known as class-III) [5,6] from mammals, slime mold or Drosophila. A consensus pattern was derived from the most conserved region in that domain.

Consensus pattern G-V-[LIVM][LIVM SEQ ID NO:4]-x(0,1)-G-x(5)-[FY]-x-[LIVM][LIVM
 SEQ ID NO:4]-[FYW]-[GS]-[DNTHKW][DNTHKW SEQ ID NO:750]-[DNT]-[IV]-
 [DNTA][DNTA SEQ ID NO:751]-x(5)-[DE]

- 5 Sequences known to belong to this class detected by the pattern ALL, except for the sea urchin *Arbacia punctulata* resact receptor which lack this domain.

Note this pattern will detect both domains of adenylate cyclases class-III.

[1] Koesling D., Boehme E., Schultz G. FASEB J. 5:2785-2791(1991).

10 [2] Garbers D.L. New Biol. 2:499-504(1990).

[3] Garbers D.L. Cell 71:1-4(1992).

[4] Yuen P.S.T., Garbers D.L. Annu. Rev. Neurosci. 15:193-225(1992).

[5] Iyengar R. FASEB J. 7:768-775(1993).

[6] Barzu O., Danchin A. Prog. Nucleic Acid Res. Mol. Biol. 49:241-283(1994).

15

956. Hemolysin-type calcium-binding region signature (HemolysinCabinD)

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These
 20 proteins, while having different functions, seem [1] to share two properties: they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which are currently known to belong to this category are:

- 25 - Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: *E. coli* (gene hlyA), *A. pleuropneumoniae* (gene appA), *A. actinomycetemcomitans* and *P. haemolytica* (leukotoxin) (gene lktA).

- 30 - Cyclolysin from *Bordetella pertussis* (gene cyaA). A multifunctional protein which is both an adenylate cyclase and a hemolysin.

- Extracellular zinc proteases: serralysin (EC 3.4.24.40) from *Serratia*, prtB and prtC from *Erwinia chrysanthemi* and aprA from *Pseudomonas aeruginosa*.

- Nodulation protein nodO from *Rhizobium leguminosarum*.

A signature pattern was derived from conserved positions in the sequence of the calcium-binding domain.

5 Consensus pattern D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D Sequences known to belong to this class detected by the pattern ALL.

Note: This pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.

10 [1] Economou A., Hamilton W.D.O., Johnston A.W.B., Downie J.A. EMBO J. 9:349-354(1990).

[2] Baumann U., Wu S., Flaherty K.M., McKay D.B. EMBO J. 12:3357-3364(1993).

957. Hint module (Hint)

15

This is an alignment of the Hint module in the Hedgehog proteins. It does not include any Inteins which also possess the Hint module.

Number of members: 36

20 [1] Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ; Medline: 97474313
"Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins." Cell 1997;91:85-97.

958. Hydantoinase/oxoprolinase (Hydantoinase)

25

This family includes the enzymes hydantoinase and oxoprolinase EC:3.5.2.9. Both reactions involve the hydrolysis of 5-membered rings via hydrolysis of their internal imide bonds [1].

Number of members: 14

30 [1] Ye GJ, Breslow EB, Meister A, Guo-jie GE\$[corrected to Ye GJ]; Medline: 97113037
"The amino acid sequence of rat kidney 5-oxo-L-prolinase determined by cDNA cloning"
[published erratum appears in J Biol Chem 1997 Feb 14;272(7):4646] J Biol Chem
1996;271:32293-32300.

959. IMP dehydrogenase / GMP reductase signature (IMPDH_N)

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase isozymes in humans [2].

GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.

IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of these regions is centered on a cysteine residue thought [3] to be involved in binding IMP. This region was used as a signature pattern.

Consensus pattern [LIVM][LIVM SEQ ID NO:4]-[RK]-[LIVM][LIVM SEQ ID NO:4]-G-[LIVM][LIVM SEQ ID NO:4]-G-x-G-S-[LIVM][LIVM SEQ ID NO:4]-C-x-T [C is the putative IMP-binding residue] Sequences known to belong to this class detected by the pattern ALL.

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem. 265:5292-5295(1990).

[3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

960. impB/mucB/samB family (IMS)

These proteins are involved in UV protection (Swiss).

Number of members: 38

961. Type II intron maturase (Intron_maturase2)

Group II introns use intron-encoded reverse transcriptase, maturase and DNA endonuclease activities for site-specific insertion into DNA [2]. Although this type of intron is self splicing in vitro they require a maturase protein for splicing in vivo. It has been shown that a specific region of the aI2 intron is needed for the maturase function [1]. This region was found to be conserved in group II introns and called domain X [3].

Number of members: 335

[1] Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P; Medline: 94301788 "Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Res 1994;22:2057-2064.

[2] Guo H, Zimmerly S, Perlman PS, Lambowitz AM; Medline: 98031910 "Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA." EMBO J 1997;16:6835-6848.

[3] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

962. LAGLIDADG endonuclease (Intron_maturase)

[1] Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL; Medline: 97331323 "The structure of I-Crel, a group I intron-encoded homing endonuclease." Nat Struct Biol 1997;4:468-476.

[2] Belfort M, Roberts RJ; Medline: 97402526 "Homing endonucleases: keeping the house in order." Nucleic Acids Res 1997;25:3379-3388.

[3] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.

Number of members: 220

963. Isopentenyl transferase (IPT)

5 Isopentenyl transferase / dimethylallyl transferase synthesizes isopentenyladenosine 5'-monophosphate, a cytokinin that induces shoot formation on host plants infected with the Ti plasmid [1].

Number of members: 16

10 [1] Canaday J, Gerad JC, Crouzet P, Otten L; Medline: 93101133 "Organization and functional analysis of three T-DNAs from the vitopine Ti plasmid pTiS4." Mol Gen Genet 1992;235:292-303.

964. Laminin EGF-like (Domains III and V) (laminin_EGF)

15 This family is like EGF but has 8 conserved cysteines instead of 6.

Number of members: 501

[1] Engel J; Medline: 93041759 "Laminins and other strange proteins." Biochemistry 1992;31:10643-10651.

20

965. Legume lectins signatures (lectin_legA)

25 Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2]. These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals).

30 Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by

800

formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.

Two signature patterns were developed specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

Consensus pattern [LIV]-[~~STAG~~][STAG SEQ ID NO:20]-V-[~~DEQV~~][DEQV SEQ ID NO:358]-[FLI]-D-[ST] [D binds manganese and calcium] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [LIV]-x-[EDQ]-[~~FYWKR~~][FYWKR SEQ ID NO:359]-V-x-[~~LIVF~~][LIVF SEQ ID NO:127]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL.

[1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

966. Malate synthase signature (malate_synthase)

Malate synthase (EC 4.1.3.2) catalyzes the aldol condensation of glyoxylate with acetyl-CoA to form malate - the second step of the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. Malate synthase is a protein of 530 to 570 amino acids whose sequence is highly conserved across species [1]. As a signature pattern, a very conserved region was selected in the central section of the enzyme.

Consensus pattern[KR]-[~~DENQ~~][DENQ SEQ ID NO:371]-H-x(2)-G-L-N-x-G-x-W-D-Y-[~~LIVM~~][LIVM SEQ ID NO:4]-F Sequences known to belong to this class detected by the pattern ALL.

[1] Bruinenberg P.G., Blaauw M., Kazemier B., Ab G. Yeast 6:245-254(1990).

801

967. MatK/TrnK amino terminal region (MatK_N)

[1] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

Number of members: 495

968. MOZ/SAS family (MOZ_SAS)

This region of these proteins has been suggested to be homologous to acetyltransferases [1]. However the similarity is not supported by standard sequence analysis.

Number of members: 15

[1] Kamine J, Elangovan B, Subramanian T, Coleman D, Chinnadurai G; Medline: 96182937 "Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator." Virology 1996;216:357-366.

[2] Reifsnyder C, Lowell J, Clarke A, Pillus L; Medline: 96376969 "Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases" [see comments] [published erratum appears in Nat Genet 1997 May;16(1):109] Nat Genet 1996;14:42-49.

969. mRNA capping enzyme (mRNA_cap_enzyme)

[1] Hakansson K, Doherty AJ, Shuman S, Wigley DB; Medline: 97304383 "X-ray crystallography reveals a large conformational change during guanylyl transfer by mRNA capping enzymes." Cell 1997;89:545-553.

Number of members: 7

970. DNA mismatch repair proteins mutS family signature (MutS_C)

Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related [2,3]. One of these families is called mutS [4,E1], it consists of:

- Prokaryotic protein mutS protein (also called hexA in *Streptococcus pneumoniae*). MutS is thought to carry out the mismatch recognition step of DNA repair.
- Eukaryotic MSH1, which is involved in mitochondrial DNA repair.
- Eukaryotic MSH2, which is involved in nuclear postreplication mismatch repair. MSH2 heterodimerizes with MSH6. In man, MSH2 is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC).
- Eukaryotic MSH3, which is probably involved in the repair of large loops.
- Eukaryotic MSH4, which is involved in meiotic recombination.
- Eukaryotic MSH5, which is involved in meiotic recombination.
- Eukaryotic MSH6 (also known as G/T mismatch binding protein), a DNA-repair protein that binds to G/T mismatches through heterodimerization with MSH2.
- Prokaryotic protein mutS2 whose function is not yet known.
- A coral (*Sarcophyton glaucum*) mitochondrial encoded mutS-like protein.

As a signature pattern for this class of mismatch repair proteins a region rich in glycine and negatively charged residues was selected. This region is found in the C-terminal section of these proteins; about 80 residues to the C-terminal of an ATP-binding site motif 'A' (P-loop) (see <PDOC00017>).

Consensus pattern[ST]-[LIVMF]-[LIVMF SEQ ID NO:2])-x-[LIVM]-[LIVM SEQ ID NO:4])-x-D-E-[LIVMFY]-[LIVMFY SEQ ID NO:18])- [GC]-[RKH]-G-[GST]- x(4)-G Sequences known to belong to this class detected by the pattern ALL, except for mutS2.

[1] Modrich P. Annu. Rev. Biochem. 56:435-466(1987).

[2] Haber L.T., Walker G.C. EMBO J. 10:2707-2715(1991).

[3] New L., Liu K., Crouse G.F. Mol. Gen. Genet. 239:97-108(1993).

[4] Eisen J.A. Nucleic Acids Res. 26:4291-4300(1998).

971. MutS family, N-terminal putative DNA binding domain (MutS_N)

This family consists of the N-terminal region of proteins in the mutS family of DNA mismatch repair proteins and is found associated with MutS_C located in the C-terminal region. The mutS family of proteins is named after the salmonella typhimurium MutS protein involved in mismatch repair; other members of the family included the eukaryotic MSH 1,2,3,4,5 and 6 proteins. These have various roles in DNA repair and recombination. Human MSH has been implicated in non-polyposis colorectal carcinoma (HNPCC) and is a mismatch binding protein [2]. The aligned region corresponds in part with domains A1, A2 (which may bind DNA) and B (which binds dsDNA in vitro) from T. thermophilus MutS as characterised in [1].

Number of members: 43

972. Domain in Myosin and Kinesin Tails (MyTH4)

Domain present twice in myosin-VIIa, and also present in 3 other myosins.

[1] Chen ZY, Hasson T, Kelley PM, Schwender BJ, Schwartz MF, Ramakrishnan M, Kimberling WJ, Mooseker MS, Corey DP; Medline: 97038686 "Molecular cloning and domain structure of human myosin-VIIa, the gene product defective in Usher syndrome 1B." Genomics 1996;36:440-448.

Number of members: 21

973. Sodium and potassium ATPases beta subunits signatures (Na_K-ATPase)

The sodium pump (Na⁺,K⁺ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane.

Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains

three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below.

```
+----+ +--+ +-----+ |||||
xxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxxCxxxxxxxxCxxxx
5  **** **<-Cyt-><TM><-----Extracellular----->
```

'C': conserved cysteine involved in a disulfide bond.

'*': position of the patterns.

- 10 Two isoforms of the beta subunit (beta-1 and beta-2) are currently known; they share about 50% sequence identity. Gastric (K⁺, H⁺) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3]; the beta chain is highly similar to the sodium pump beta subunits. Two signature patterns were developed for beta subunits. The first is located in the cytoplasmic domain, while the second is found in the extracellular
- 15 domain and contains two of the cysteines involved in disulfide bonds.

Consensus pattern [FYW]-x(2)-[FYW]-x-[FYW]-[DN]-x(6)-[LIVM][LIVM SEQ ID NO:4)]-G-R-T-x(3)-W Sequences known to belong to this class detected by the pattern ALL.

- 20 Consensus pattern [RK]-x(2)-C-[RKQWI][RKQWI SEQ ID NO:752)]-x(5)-L-x(2)-C-[SA]-G [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for the beta subunit of the sodium pump of brine shrimp whose sequence is highly divergent in that region.

- 25 [1] Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. Annu. Rev. Physiol. 53:565-584(1991).
[2] McDonough A.A., Gerring K., Farley R.A. FASEB J. 4:1598-1605(1990).
[3] Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T.,
- 30 Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. Proc. Natl. Acad. Sci. U.S.A. 87:6418-6422(1990).

974. Respiratory-chain NADH dehydrogenase subunit 1 signatures (NADHdh)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there are fifteen which are located in the membrane part, seven of which are encoded by the mitochondrial and chloroplast genomes of most species. The most conserved of these organelle-encoded subunits is known as subunit 1 (gene ND1 in mitochondrion, and NDH1 in chloroplast) and seems to contain the ubiquinone binding site.

The ND1 subunit is highly similar to subunit 4 of *Escherichia coli* formate hydrogenlyase (gene hycD), subunit C of hydrogenase-4 (gene hyfC). *Paracoccus denitrificans* NQO8 and *Escherichia coli* nuoH NADH-ubiquinone oxidoreductase subunits also belong to this family [3]. Two signature patterns were developed based on conserved regions of this subunit.

Consensus pattern G-[LIVMFYKRS][LIVMFYKRS SEQ ID NO:753]-
[LIVMAGP][LIVMAGP SEQ ID NO:415]-Q-x-[LIVMFY][LIVMFY SEQ ID NO:18]-x-
D-[AGIM][AGIM SEQ ID NO:754]-[LIVMFTA][LIVMFTA SEQ ID NO:386]-K-
[LVMYST][LVMYST SEQ ID NO:755]-[LIVMFYG][LIVMFYG SEQ ID NO:168]-x-

[KR]-[EQG] Sequences known to belong to this class detected by the pattern ALL, except for watermelon and *Leishmania* ND1.

Consensus pattern P-F-D-[LIVMFYQ][LIVMFYQ SEQ ID NO:188]-
[STAGPVM][STAGPVM SEQ ID NO:756]-E-[GAC]-E-x-[EQ]-[LIVMS][LIVMS SEQ ID
NO:429]-x(2)-G Sequences known to belong to this class detected by the pattern ALL,
except for *Chlamydomonas reinhardtii* and *Pisaster ochraceus* ND1, and tobacco NDH1.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

975. Nickel-dependent hydrogenases large subunit signatures (NiFeSe_Hases)

Hydrogenases are enzymes that catalyze the reversible activation of hydrogen and which occur widely in prokaryotes as well as in some eukaryotes. There are various types of hydrogenases, but all of them seem to contain at least one iron-sulfur cluster. They can be broadly divided into two groups: hydrogenases containing nickel and, in some cases, also selenium (the [NiFe] and [NiFeSe] hydrogenases) and those lacking nickel (the [Fe] hydrogenases).

The [NiFe] and [NiFeSe] hydrogenases are heterodimer that consist of a small subunit that contains a signal peptide and a large subunit. All the known large subunits seem to be evolutionary related [1]; they contain two Cys-x-x- Cys motifs; one at their N-terminal end; the other at their C-terminal end. These four cysteines are involved in the binding of nickel [2]. In the [NiFeSe] hydrogenases the first cysteine of the C-terminal motif is a selenocysteine which has experimentally been shown to be a nickel ligand [3]. Two patterns were developed which are centered on the Cys-x-x-Cys motifs.

Alcaligenes eutrophus possess a NAD-reducing cytoplasmic hydrogenase (hoxS) [4]; this enzyme is composed of four subunits. Two of these subunits (beta and delta) are responsible for the hydrogenase reaction and are evolutionary related to the large and small subunits of membrane-bound hydrogenases. The alpha subunit of coenzyme F420 hydrogenase (EC 1.12.99.1) (FRH) from archaeobacterial methanogens also belongs to this family.

Consensus pattern R-G-[LIVMF][LIVMF SEQ ID NO:2]-E-x(15)-[QESM][QESM SEQ ID NO:757]-R-x-C-G-[LIVM][LIVM SEQ ID NO:4]-C [The two C's are nickel ligands]

Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FY]-D-P-C-[LIM]-[ASG]-C-x(2,3)-H [The two C's are nickel ligands]
Sequences known to belong to this class detected by the pattern ALL.

[1] Menon N.K., Robbins J., Peck H.D. Jr., Chatelus C.Y., Choi E.-S., Przybyla A.E. J. Bacteriol. 172:1969-1977(1990).

[2] Volbeda A., Charon M.-H., Piras C., Hatchikian E.C., Frey M., Fontecilla-Camps J.C. Nature 373:580-587(1995).

[3] Eidsness M.K., Scott R.A., Prickrill B., der Vartanian D.V., LeGall J., Moura I., Moura J.J.G., Peck H.D. Jr. Proc. Natl. Acad. Sci. U.S.A. 86:147-151(1989).

[4] Tran-Betcke A., Warnecke U., Boecker C., Zaborosch C., Friedrich B. J. Bacteriol. 172:2920-2929(1990).

5

976. NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus (oxidored_q1_C)

This sub-family represents a carboxyl terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 from chloroplasts are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

10

Number of members: 572

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

15

977. NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus (oxidored_q1_N)

This sub-family represents an amino terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

20

Number of members: 546

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

25

978. oxidored_q2. NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4L (EC 1.6.5.3). ND4L OR NAD4L. Arabidopsis thaliana (Mouse-ear cress). Mitochondrion. OC Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis. CATALYTIC ACTIVITY: NADH + UBIQUINONE = NAD(+) + UBIQUINOL.

30

[1] SEQUENCE FROM N.A. MEDLINE; 93156682. Brandt P., Sunkel S., Unseld M., Brennicke A., Knoop V.; "The nad4L gene is encoded between exon c of nad5 and orf25 in the Arabidopsis mitochondrial genome."; Mol. Gen. Genet. 236:33-38(1992).

[2] SEQUENCE FROM N.A. STRAIN=CV. COLUMBIA; MEDLINE; 97141919 Unseld M., Marienfeld J.R., Brandt P., Brennicke A.; "The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides."; Nat. Genet. 15:57-61(1997).

979. oxidored_q4. Protein name NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN 3, CHLOROPLAST. Synonym(s)EC 1.6.5.3. Gene name(s)NDHC OR NDH3 From Zea mays (Maize) Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Zea.

CATALYTIC ACTIVITY: NADH + PLASTOQUINONE = NAD(+) + PLASTOQUINOL.

SIMILARITY: BELONGS TO THE COMPLEX I SUBUNIT 3 FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 89281491. Steinmueller K., Ley A.C., Steinmetz A.A., Sayre R.T., Bogorad L.; "Characterization of the ndhC-psbG-ORF157/159 operon of maize plastid DNA and of the cyanobacterium Synechocystis sp. PCC6803."; Mol. Gen. Genet. 216:60-69(1989).

[2] SEQUENCE FROM N.A. MEDLINE; 95395841. Maier R.M., Neckermann K., Igloi G.L., Koessel H.; "Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing."; J. Mol. Biol. 251:614-628(1995).

980. PAC: PAC motif

PAC motif occurs C-terminal to a subset of all known PAS motifs. It is proposed to contribute to the PAS domain fold [3]. Number of members: 181

[1] Medline: 97446881 PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox. Zhulin IB, Taylor BL, Dixon R; Trends Biochem Sci 1997;22:331-333.

[2] Medline: 95275818. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Borgstahl GE, Williams DR, Getzoff ED; Biochemistry 1995;34:6278-6287.

[3] Medline: 98044337. PAS: a multifunctional domain family comes to light. Ponting CP, Aravind L; Curr Biol 1997;7:674-677.

981. PARP: Poly(ADP-ribose) polymerase catalytic region.

- 5 Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD⁺ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Poly(ADP-ribose) polymerase is a regulatory component induced by DNA damage.
- 10 The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19

[1] Medline: 96353841 Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Proc Natl Acad Sci U S A 1996;93:7481-7485.

[2] Medline: 93293867 The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale purification, and characterization. Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G, Althaus FR; J Biol Chem 1993;268:13454-13461.

982. PC_rep: Proteasome/cyclosome repeat

- [1] Medline: 97348748 A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). Lupas A, Baumeister W, Hofmann K; Trends Biochem Sci 1997;22:195-196.
- Number of members: 112

983. Peptidase_M1: Peptidase family M1

- Members of this family are aminopeptidases. The members differ widely in specificity, hydrolysing acidic, basic or neutral N-terminal residues. This family includes leukotriene-A4 hydrolase Swiss:P09960, this enzyme also has an aminopeptidase activity [1]. Number of members: 72

[1] Medline: 95405261 Evolutionary families of metallopeptidases. Rawlings ND, Barrett AJ; Meth Enzymol 1995;248:183-228.

984. Neutral zinc metallopeptidases, zinc-binding region signature (Peptidase_M8)

5 PROSITE cross-reference(s) PS00142; ZINC_PROTEASE

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a
10 superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).
- 15 - Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- Yeast alanine/arginine aminopeptidase (gene AAP1).
- 20 - Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

Family M2

- 25 - Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic
30 degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).

- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).
- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).
- Yeast hypothetical protein YKL134c.

Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus.
- Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB).
- Extracellular elastase from Staphylococcus epidermidis.
- Extracellular protease prt1 from Erwinia carotovora.
- Extracellular minor protease smp from Serratia marcescens.
- Vibriolysin (EC 3.4.24.25) from various species of Vibrio.
- Protease prtA from Listeria monocytogenes.
- Extracellular proteinase proA from Legionella pneumophila.

Family M5

- Mycolysin (EC 3.4.24.31) from Streptomyces cacaoi.

Family M6

- Immune inhibitor A from Bacillus thuringiensis (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- Streptomyces extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of Leishmania.

Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

Family M10A

- 5 - Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.
- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene *aprA*).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.
- Yeast hypothetical protein YIL108w.

10 Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- 15 - Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

20

Family M11

- *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

Family M12A

- 25 - Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog
- 30 of BMP-1 is the dorsal-ventral patterning protein *tolloid*.
- Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN from *Strongylocentrotus purpuratus*.
- *Caenorhabditis elegans* protein *toh-2*.

- *Caenorhabditis elegans* hypothetical protein F42A10.8.
- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimereylisin I (EC 3.4.25.52) and II (EC 3.4.25.53).
- Mouse cell surface antigen MS2.

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

5

Family M35

- Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

10

Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

20

Consensus pattern[GSTALIVN][GSTALIVN SEQ ID NO:679]-x(2)-H-E-
[LIVMFYW][LIVMFYW SEQ ID NO:26]-[DEHRKP+][DEHRKP SEQ ID NO:680])-H-x-
[LIVMFYWGSPQ][LIVMFYWGSPQ SEQ ID NO:681]

[The two H's are zinc ligands] [E is the active site residue]

25

Sequences known to belong to this class detected by the patternALL, except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT57; including *Neurospora crassa* conidiation-specific protein 13 which could be a zinc-protease.

[1]Jongeneel C.V., Bouvier J., Bairoch A. FEBS Lett. 242:211-214(1989).

30

[2]Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).

[4]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

- [5]Woessner J. Jr. FASEB J. 5:2145-2154(1991).
 [6]Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373:381-385(1992).
 [7]Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993).
 [8]Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4:179-185(1994).

5

985. PHO4: Phosphate transporter family

This family includes PHO-4 from *Neurospora crassa* which is a Na(+)-phosphate symporter [1]. This family also contains the leukemia virus receptor Swiss:Q08344. Number of members: 41

10

[1] Medline: 95249577 Repressible cation-phosphate symporters in *Neurospora crassa*. Versaw WK, Metzenberg RL; Proc Natl Acad Sci U S A 1995;92:3884-3887.

986. Photosynthetic reaction center proteins signature (photoRC)

15

PROSITE cross-reference(s): PS00244; REACTION_CENTER

In the photosynthetic reaction center of purple bacteria, two homologous integral membrane proteins, L(ight) and M(edium), are known to be essential to the light-mediated water-splitting process. In the photosystem II of eukaryotic chloroplasts two related proteins are involved: the D1 (psbA) and D2 proteins (psbD). These four types of protein probably evolved from a common ancestor [see 1,2 for recent reviews].

20

A signature pattern was developed which include two conserved histidine residues. In L and M chains, the first histidine is a ligand of the magnesium ion of the special pair bacteriochlorophyll, the second is a ligand of a ferrous non-heme iron atom. In photosystem II these two histidines are thought to play a similar role.

25

Consensus pattern[NQH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC][SAGC SEQ ID NO:758]-x-H-[SAG](2)

30

[The first H is a magnesium ligand] [The second H is a iron ligand]

Sequences known to belong to this class detected by the patternALL, except for broad bean psbA which has Gln instead of the second His.

[1] Michel H., Deisenhofer J. *Biochemistry* 27:1-7(1988).

[2] Barber J. *Trends Biochem. Sci.* 12:321-326(1987).

987. phytochrome: Phytochrome region

5 This family contains a region specific to phytochrome proteins. Number of members:
145

988. PI3K_C2: C2 domain

Phosphoinositide 3-kinase region postulated to contain a C2 domain. Outlier of C2 family.

10 Number of members: 39

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; *FEBS Lett* 1997;410:91-95.

[2] Medline: 97398940 Phosphoinositide 3-kinases: a conserved family of signal transducers.
15 Vanhaesebroeck B, Leever SJ, Panayotou G, Waterfield MD; *Trends Biochem Sci*
1997;22:267-272.

989. PI3Ka: Phosphoinositide 3-kinase family, accessory domain (PIK domain)

PIK domain is conserved in all PI3 and PI4-kinases. Its role is unclear but it has been
20 suggested [2] to be involved in substrate presentation.

Number of members: 47

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; *FEBS Lett* 1997;410:91-95.

25 [2] Medline: 94069320 Phosphatidylinositol 4-kinase: gene structure and requirement for
yeast cell viability. Flanagan CA, Schnieders EA, Emerick AW, Kunisawa R, Admon A,
Thorner J; *Science* 1993;262:1444-1448.

990. P-II protein signatures

30 PROSITE cross-reference(s): PS00496; PII_GLNB_UMP, PS00638; PII_GLNB_CTER

The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-

ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme. Conversely, in nitrogen excess, P-II-UMP is deuridylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (*glnA*) by preventing NR-II (*ntrB*) to phosphorylate NR-I (*ntrC*) which is the transcriptional activator of *glnA*. Once P-II is uridylylated, these events are reversed.

P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is urydylated is located in the central part of the protein.

In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being urydylated.

In methanogenic archaeobacteria, the nitrogenase iron protein gene (*nifH*) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.

In the red alga, *Porphyra purpurea*, there is a *glnB* homolog encoded in the chloroplast genome.

Other proteins highly similar to *glnB* are:

- *Bacillus subtilis* protein *nrgB* [5].
- *Escherichia coli* hypothetical protein *ybaI* [6].

Two signature patterns were developed for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the urydylated tyrosine, the other is derived from a conserved region in the C-terminal part of the P-II protein.

Consensus pattern Y-[KR]-G-[AS]-[AE]-Y [The second Y is uridylated]
Sequences known to belong to this class detected by the pattern ALL *glnB*'s from eubacteria.

Consensus pattern[ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM][LIVM SEQ ID NO:4]-
x(2)-[LIVM][LIVM SEQ ID NO:4]

[1]Magasanik B. Biochimie 71:1005-1012(1989).

5 [2]Holtel A., Merrick M. Mol. Gen. Genet. 215:134-138(1988).

[3]Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure
2:981-990(1994).

[4]Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).

[5]Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).

10 [6]Allikmets R., Gerrard B.C., Court D., Dean M.C. Gene 136:231-236(1993).

991. PIP5K: Phosphatidylinositol-4-phosphate 5-Kinase

This family contains a region from the common kinase core found in the type I
phosphatidylinositol-4-phosphate 5-kinase (PIP5K) family as described in [1]. The family
15 consists of various type I, II and III PIP5K enzymes. PIP5K catalyses the formation of
phosphoinositol-4,5-bisphosphate via the phosphorylation of phosphatidylinositol-4-
phosphate a precursor in the phosphoinositide signaling pathway. Number of members: 33

[1] Medline: 98204859. Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the
20 third isoform and deletion/substitution analysis of members of this novel lipid kinase family.
Ishihara H, Shibasaki Y, Kizuki N, Wada T, Yazaki Y, Asano T, Oka Y; J Biol Chem
1998;273:8741-8748.

[2] Medline: 97115834 Type I phosphatidylinositol-4-phosphate 5-kinases are distinct
members of this novel lipid kinase family. Loijens JC, Anderson RA; J Biol Chem 1996
25 20;271:32937-32943.

992. PolyA_pol: Poly A polymerase family

This family includes nucleic acid independent RNA polymerases, such as Poly(A)
polymerase, which adds the poly (A) tail to mRNA EC:2.7.7.19. This family also includes the
30 tRNA nucleotidyltransferase that adds the CCA to the 3' of the tRNA
EC:2.7.7.25. Number of members: 31

[1] Medline: 93066242 Identification of the gene for an Escherichia coli poly(A) polymerase. Cao GJ, Sarkar N; Proc Natl Acad Sci U S A 1992;89:10380-10384.

993. Photosystem I psaA and psaB proteins signature (psaA_psaB)

5 PROSITE cross-reference(s)PS00419; PHOTOSYSTEM_I_PSAAB

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin. PSI is found in the chloroplast of plants and cyanobacteria. The electron transfer components of the reaction center of
10 PSI are a primary electron donor P-700 (chlorophyll dimer) and five electron acceptors: A0 (chlorophyll), A1 (a phylloquinone) and three 4Fe-4S iron-sulfur centers: Fx, Fa, and Fb.

PsaA and psaB, two closely related proteins, are involved in the binding of P700, A0, A1, and Fx. psaA and psaB are both integral membrane proteins of 730 to 750 amino acids that
15 seem to contain 11 transmembrane segments. The Fx 4Fe-4S iron-sulfur center is bound by four cysteines; two of these cysteines are provided by the psaA protein and the two others by psaB. The two cysteines in both proteins are proximal and located in a loop between the ninth and tenth transmembrane segments. A leucine zipper motif seems to be present [2] downstream of the cysteines and could contribute to dimerization of psaA/psaB.

20

The signature pattern for these proteins is based on the perfectly conserved region that includes the two iron-sulfur binding cysteines.

Consensus patternC-D-G-P-G-R-G-G-T-C [The two C's bind the iron-sulfur center]

25 [1]Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).

[2]Webber A.N., Malkin R. FEBS Lett. 264:1-14(1990).

994. PSBH: Photosystem II 10 kDa phosphoprotein

This protein is phosphorylated in a light dependent reaction.

30 Number of members: 20

995. PsbJ

This family consists of the photosystem II reaction center protein PsbJ from plants and Cyanobacteria. In *Synechocystis* sp. PCC 6803 PsbJ regulates the number of photosystem II centers in thylakoid membranes, it is a predicted 4kDa protein with one membrane spanning domain [1]. Number of members: 20

5

[1] Medline: 93131892. Genetic and immunological analyses of the cyanobacterium *Synechocystis* sp. PCC 6803 show that the protein encoded by the *psbJ* gene regulates the number of photosystem II centers in thylakoid membranes. Lind LK, Shukla VK, Nyhus KJ, Pakrasi HB; J Biol Chem 1993;268:1575-1579.

10

996. PSBT: Photosystem II reaction centre T protein

The exact function of this protein is unknown. It probably consists of a single transmembrane spanning helix. The Swiss:P37256 protein, appears to be (i) a novel photosystem II subunit and (ii) required for maintaining optimal photosystem II activity under adverse growth conditions [1]. Number of members: 17

15

[1] Medline: 94298765. The chloroplast *ycf8* open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. Monod C, Takahashi Y, Goldschmidt-Clermont M, Rochaix JD; EMBO J 1994;13:2747-2754.

20

997. PSI_8. PHOTOSYSTEM I REACTION CENTRE SUBUNIT VIII. Synonym(s)PSI-I. Gene name(s)PSAI. From *Hordeum vulgare* (Barley). Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Hordeum.

25

MAY HELP IN THE ORGANIZATION OF THE PSAL SUBUNIT. BELONGS TO THE PSAL FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 90036933. Scheller H.V., Okkels J.S., Hoej P.B., Svendsen I., Roepstorff P., Moeller B.L.; "The primary structure of a 4.0-kDa photosystem I polypeptide encoded by the chloroplast *psaI* gene."; J. Biol. Chem. 264:18402-18406(1989).

30

998. PSI_PsaJ: Photosystem I reaction centre subunit IX / PsaJ

This family consists of the photosystem I reaction centre subunit IX or PsaJ from various organisms including *Synechocystis* sp. (strain pcc 6803), *Pinus thunbergii* (green pine) and *Zea mays* (maize). PsaJ Swiss:P19443 is a small 4.4kDa, chloroplastal encoded, hydrophobic subunit of the photosystem I reaction complex its function is not yet fully understood [1].

- 5 PsaJ can be cross-linked to PsaF Swiss:P12356 and has a single predicted transmembrane domain it has a proposed role in maintaining PsaF in the correct orientation to allow for fast electron transfer from soluble donor proteins to P700+ [1]. Number of members: 18

[1] Medline: 99238330. A large fraction of PsaF is nonfunctional in photosystem I complexes lacking the PsaJ subunit. Fischer N, Boudreau E, Hippler M, Drepper F, Haehnel W, Rochaix JD; Biochemistry 1999;38:5546-5552.

[2] Medline: 93252282. Genes encoding eleven subunits of photosystem I from the thermophilic cyanobacterium *Synechococcus* sp. Muhlenhoff U, Haehnel W, Witt H, Herrmann RG; Gene 1993;127:71-78.

15

999. PSII. Protein namePHOTOSYSTEM II P680 CHLOROPHYLL A APOPROTEIN. Synonym(s)CP-47 PROTEIN. Gene name(s)PSBB. From *Hordeum vulgare* (Barley), Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; *Hordeum*.

20

FUNCTION: THIS PROTEIN CONJUGATES WITH CHLOROPHYLL & CATALYZES THE PRIMARY LIGHT-INDUCED PHOTOCHEMICAL PROCESSES OF PHOTOSYSTEM II. SUBCELLULAR LOCATION: CHLOROPLAST THYLAKOID MEMBRANE. SIMILARITY: BELONGS TO THE PSBB / PSBC FAMILY.

25

[1] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 89240047. Andreeva A.V., Buryakova A.A., Reverdatto S.V., Chakhmakhcheva O.G., Efimov V.A.; "Nucleotide sequence of the 5.2 kbp barley chloroplast DNA fragment, containing psbB-psbH-petB-petD gene cluster."; Nucleic Acids Res. 17:2859-2860(1989).

30

[2] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 92207253. Efimov V.A., Andreeva A.V., Reverdatto S.V., Chakhmakhcheva O.G.; "Photosystem II of rye. Nucleotide sequence of the psbB, psbC, psbE, psbF, psbH genes of rye and chloroplast DNA regions adjacent to them."; Bioorg. Khim. 17:1369-1385(1991).

[3] SEQUENCE OF 411-420. Hinz U.G.; "Isolation of the photosystem II reaction center complex from barley. Characterization by circular dichroism spectroscopy and amino acid sequencing."; Carlsberg Res. Commun. 50:285-298(1985).

5 1000. QRPTase. Quinolate phosphoribosyl transferase.

Quinolate phosphoribosyl transferase (QPRTase) or nicotinate-nucleotide pyrophosphorylase EC:2.4.2.19 is involved in the de novo synthesis of NAD in both prokaryotes and eukaryotes. It catalyses the reaction of quinolinic acid with 5-phosphoribosyl-1-pyrophosphate (PRPP) in the presence of Mg^{2+} to give rise to nicotinic acid mononucleotide (NaMN), pyrophosphate and carbon dioxide [1,2]. Number of members: 26.

[1]Medline: 97169443. A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. Eads JC, Ozturk D, Wexler TB, Grubmeyer C, Sacchettini JC; Structure 1997;5:47-58.

[2]Medline: 96139309. The sequencing expression, purification, and steady-state kinetic analysis of quinolate phosphoribosyl transferase from Escherichia coli. Bhatia R, Calvo KC; Arch Biochem Biophys 1996;325:270-278.

20 1001. R3H domain

The name of the R3H domain comes from the characteristic spacing of the most conserved arginine and histidine residues. The function of the domain is predicted to be binding ssDNA. Number of members: 28

[1]Medline: 99003905 The R3H motif: a domain that binds single-stranded nucleic acids. Grishin NV; Trends Biochem Sci 1998;23:329-330.

1002. recF protein signatures (RecF)

30 The prokaryotic protein recF [1,2] is a single-stranded DNA-binding protein which also probably binds ATP. RecF is involved in DNA metabolism; it is required for recombinational DNA repair and for induction of the SOS response. RecF is a protein of about 350 to 370 amino acid residues; there is a conserved ATP-binding site motif 'A' (P-loop) in the N-

terminal section of the protein as well as two other conserved regions, one located in the central section, and the other in the C-terminal section. Signature patterns were derived from these two regions.

5 Consensus pattern [LIVM][LIVM SEQ ID NO:4]-x(4)-[LIF]-x(6)-[LIF]-[LVF]-x-[GE]-
[GSTAD][GSTAD SEQ ID NO:759]-[PA]- x(2)-R-R-x-[FYW]-[LIVMF][LIVMF SEQ ID
NO:2]-D Sequences known to belong to this class detected by the pattern ALL.

10 Consensus pattern[LIVMFY][LIVMFY SEQ ID NO:18](2)-x-D-x(2,3)-[SA]-[EH]-L-D-
x(2)-[KRH]-x(3)-L Sequences known to belong to this class detected by the patternALL,
except for T. palidum recF.

[1] Sandler S.J., Chackerian B., Li J.T., Clark A.J. Nucleic Acids Res. 20:839-845(1992).

[2] Alonso J.C., Fisher L.M.; Mol. Gen. Genet. 246:680-686(1995).

15

1003. RibD C-terminal domain (RibD_C)

The function of this domain is not known, but it is thought to be involved in riboflavin biosynthesis. This domain is found in the C terminus of RibD/RibG Swiss:P25539, in
20 combination with dCMP_cyt_deam, as well as in isolation in some archaebacterial proteins Swiss:P95872.

Number of members: 21

1004. Ribosomal protein L16 signatures (Ribosomal_L16)

25

Ribosomal protein L16 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L16 is known to bind directly the 23S rRNA and to be located at the A site of the peptidyltransferase center. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

30

- Eubacterial L16.
- Algal and plant chloroplast L16.
- Cyanelle L16.
- Plant mitochondrial L16.

L16 is a protein of 133 to 185 amino-acid residues. As signature patterns, we selected two conserved regions in the central section of these proteins.

Consensus pattern [KR](2)-x-[GSAC][GSAC SEQ ID NO:93]-[KRQVA][KROVA SEQ ID NO:760]-[LIVM][LIVM SEQ ID NO:4]-W-[LIVM][LIVM SEQ ID NO:4]-[KR]-[LIVM][LIVM SEQ ID NO:4]-[LFY]-[AP] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern R-M-G-x-[GR]-K-G-x(4)-[FWKR][FWKR SEQ ID NO:761] Sequences known to belong to this class detected by the pattern ALL.

[1] Otake E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

1005. Ribosomal protein L32e signature (Ribosomal_L32E)

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L32 [1].
- Drosophila RP49 [2].
- Trichoderma harzianum L32 [3].
- Yeast L32e (YBL092w).
- Archaebacterial L32e [4].

These proteins have 135 to 240 amino-acid residues. As a signature pattern, a stretch of about 20 residues located in the N-terminal part of these proteins was selected.

Consensus pattern F-x-R-x(4)-[KR]-x(2)-[KR]-[LIVMF][LIVMF SEQ ID NO:2]-x(3,5)-W-R-[KR]-x(2)-G Sequences known to belong to this class detected by the pattern ALL.

[1] Jacks C.M., Powaser C.B., Hackett P.B. Gene 74:565-570(1988).

[2] Aguade M. Mol. Biol. Evol. 5:433-441(1988).

[3] Lora J.M., Garcia I., Benitez T., Llobell A., Pintor-Toro J.A. Nucleic Acids Res. 21:3319-3319(1993).

[4] Arndt E., Scholzen T., Kroemer W., Hatakeyama T., Kimura M. Biochimie 73:657-668(1991).

1006. (Ribosomal_S3) Ribosomal protein S3 signature

5 PROSITE: PDOC00474. PROSITE cross-reference(s) PS00548; RIBOSOMAL_S3

Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In *Escherichia coli*, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- 10 -Eubacterial S3.
- Algal and plant chloroplast S3.
- Cyanelle S3.
- Archaeobacterial S3.
- Plant mitochondrial S3.
- 15 -Vertebrate S3.
- Insect S3.
- Caenorhabditis elegans S3 (C23G10.3).
- Yeast S3 (Rp13).

S3 is a protein of 209 to 559 amino-acid residues. A conserved region located in the C-terminal section was selected as a signature pattern.

Consensus pattern[GSTA][GSTA SEQ ID NO:19]-[KR]-x(6)-G-x-[LIVMT][LIVMT SEQ ID NO:1]-x(2)-[NQSCH][NQSCH SEQ ID NO:519]-x(1,3)-[LIVFCA][LIVFCA SEQ ID NO:520]-x(3)-[LIV]-[DENQ][DENQ SEQ ID NO:371]-x(7)-[LMT]-x(2)-G-x(2)-[GS].

25 Sequences known to belong to this class detected by the patternALL, except for some mitochondrial S3.

[1]Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

30 1007. RimM - RimM

The RimM protein is essential for efficient processing of 16S rRNA [1]. The RimM protein was shown to have affinity for free ribosomal 30S subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14.

[1]Medline: 98083058. RimM and RbfA are essential for efficient processing of 16S rRNA in *Escherichia coli*. Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; J Bacteriol 1998;180:73-82.

5

1008. RNA_pol_A - RNA polymerase alpha subunit

-!- RNA polymerases catalyse the DNA dependent polymerisation of RNA. Prokaryotes contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial and chloroplast polymerases).

10

-!- Members of this family include: A subunit from eukaryotes, gamma subunit from cyanobacteria, beta' subunit from eubacteria, A' subunit from archaeobacteria, B'' from chloroplasts. Number of members: 139.

[1]Medline: 97066998. Structural modules of the large subunits of RNA polymerase.

15

Introducing archaeobacterial and chloroplast split sites in the beta and beta' subunits of *Escherichia coli* RNA polymerase. Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst SA, Goldfarb A; J Biol Chem 1996;271:27969-27974.

1009. RuBisCO_large - Ribulose biphosphate carboxylase large chain active site

20

PROSITE: PDOC00142; PROSITE cross-reference(s) PS00157; RUBISCO_LARGE

Ribulose biphosphate carboxylase (EC 4.1.1.39) (RuBisCO) [1,2] catalyzes the initial step in Calvin's reductive pentose phosphate cycle in plants as well as purple and green bacteria. It consists of a large catalytic unit and a small subunit of undetermined function. In plants, the large subunit is coded by the chloroplastic genome while the small subunit is encoded in the nuclear genome. Molecular activation of RuBisCO by CO₂ involves the formation of a carbamate with the epsilon-amino group of a conserved lysine residue. This carbamate is stabilized by a magnesium ion. One of the ligands of the magnesium ion is an aspartic acid residue close to the active site lysine [3]. A pattern was developed which includes both the active site residue and the metal ligand, and which is specific to RuBisCO large chains.

30

Consensus pattern G-x-[DN]-F-x-K-x-D-E [K is the active site residue] [The second D is a magnesium ligand]. Sequences known to belong to this class detected by the pattern ALL, except for *Cheilopleuria bicuspidis* RuBisCO.

- 5 [1] Mizioro H.M., Lorimer G.H. Annu. Rev. Biochem. 52:507-535(1983).
- [2] Akazawa T., Takabe T., Kobayashi H. Trends Biochem. Sci. 9:380-383(1984).
- [3] Andersson I., Knight S., Schneider G., Lindqvist Y., Lundqvist T., Branden C.-I., Lorimer G.H. Nature 337:229-234(1989).

10 1010. Rve - Integrase core domain

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain Integrase_Zn. This domain is the central catalytic domain. The carboxyl terminal domain that is a non-specific DNA binding domain integrase. The catalytic domain acts as an endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended viral DNA made by reverse transcription. This domain also catalyses the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Number of members: 694.

- 20 [1] Medline: 95099322. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Science 1994;266:1981-1986.

25 1011. (SBP_bac_3) Bacterial extracellular solute-binding proteins, family 3 signature PROSITE: PDOC00798. PROSITE cross-reference(s) PS01039; SBP_BACTERIAL_3

Bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-binding proteins (ABC transporters; see <PDOC00185>) and a high affinity periplasmic solute-binding protein. The latter are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute through the membrane by binding to external sites of the integral membrane proteins of the efflux system.

In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be grouped [1] into eight families of clusters, which generally correlate with the nature of the solute bound.

Family 3 groups together specific amino acids and opine-binding periplasmic proteins and a periplasmic homolog with catalytic activity:

-Histidine-binding protein (gene *hisJ*) of *Escherichia coli* and related bacteria. An

homologous lipoprotein exists in *Neisseria gonorrhoeae*.

-Lysine/arginine/ornithine-binding proteins (LAO) (gene *argT*) of *Escherichia coli* and related bacteria are involved in the same transport system than *hisJ*. Both solute-binding proteins interact with a common membrane-bound receptor *hisP* of the binding protein dependent transport system HisQMP.

-Glutamine-binding proteins (gene *glnH*) of *Escherichia coli* and *Bacillus stearothermophilus*.

-Glutamate-binding protein (gene *gluB*) of *Corynebacterium glutamicum*.

-Arginine-binding proteins *artI* and *artJ* of *Escherichia coli*.

-Nopaline-binding protein (gene *nocT*) from *Agrobacterium tumefaciens*.

-Octopine-binding protein (gene *occT*) from *Agrobacterium tumefaciens*.

-Major cell-binding factor (CBF1) (gene: *peb1A*) from *Campylobacter jejuni*.

-*Bacteroides nodosus* protein *aabA*.

-Cyclohexadienyl/arogenate dehydratase of *Pseudomonas aeruginosa*, a periplasmic enzyme which forms an alternative pathway for phenylalanine biosynthesis.

-*Escherichia coli* protein *fliY*.

-*Vibrio harveyi* protein *patH*.

-*Escherichia coli* hypothetical protein *ydhW*.

-*Bacillus subtilis* hypothetical protein *yckB*.

-Bacillus subtilis hypothetical protein yckK.

The signature pattern is located near the N-terminus of the mature proteins.

Consensus pattern G-[FYIL][FYIL SEQ ID NO:644)]-[DE]-[LIVMT][LIVMT SEQ ID
 5 NO:1)]-[DE]-[LIVMF][LIVMF SEQ ID NO:2)]-x(3)-[LIVMA][LIVMA SEQ ID NO:30)]-
 [VAGC][VAGC SEQ ID NO:762)]-x(2)-[LIVMAGN][LIVMAGN SEQ ID NO:763)]

Sequences known to belong to this class detected by the pattern ALL.

[1] Tam R., Saier M.H. Jr. Microbiol. Rev. 57:320-346(1993).

10

1012. Sec7 - Sec7 domain

The Sec7 domain is a guanine-nucleotide-exchange-factor (GEF) for the arf family [2].

Number of members: 32.

15

[1] Medline: 98169075. Structure of the Sec7 domain of the Arf exchange factor. ARNO. Cherfils J, Menetrey J, Mathieu M, Le Bras G, Robineau S, Beraud-Dufour S, Antonny B, Chardin P; Nature 1998;392:101-105.

[2] Medline: 97100951. A human exchange factor for ARF contains Sec7- and pleckstrin-homology domains. Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Jackson
 20 CL, Chabre M. Nature 1996;384:481-484.

1013. SecA_protein. SecA protein, amino terminal region

SecA protein binds to the plasma membrane where it interacts with proOmpA to support translocation of proOmpA through the membrane. SecA protein achieves this translocation,
 25 in association with SecY protein, in an ATP dependent manner. SecA possesses the ATPase activity. The carboxyl terminus has similarity with the helicase carboxyl terminus. See Ribosomal_L5. Number of members: 45.

25

[1] Medline: 98309858. Amino-terminal region of SecA is involved in the function of SecG
 30 for protein translocation into Escherichia coli membrane vesicles. Mori H, Sugiyama H, Yamanaka M, Sato K, Tagaya M, Mizushima S; J Biochem (Tokyo) 1998;124:122-129.

[2]Medline: 89251629. SecA protein hydrolyzes ATP and is an essential component of the protein translocation ATPase of Escherichia coli. Lill R, Cunningham K, Brundage LA, Ito K, Oliver D, Wickner W; EMBO J 1989;8:961-966.

5 1014. Seedstore_2S - 2S seed storage family

Members of this family are composed of two chains (both included in the alignment), these are co-translated and later cleaved. The two chains are disulphide linked together. Number of members: 27.

10 [1]Medline: 97121264. 1H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R; Biochemistry 1996;35:15672-15682.

1015. Smr - Smr domain

15 This family includes the Smr (Small MutS Related) proteins, and the C-terminal region of the MutS2 protein. It has been suggested that this domain interacts with the MutS1 Swiss:P23909 protein in the case of Smr proteins and with the N-terminal MutS related region of MutS2 Swiss:P94545 [1]. Number of members: 14.

20 [1]Medline: 10431172. Smr: a bacterial and eukaryotic homologue of the C-terminal region of the MutS2 family. Moreira D, Philippe H; Trends Biochem Sci 1999;24:298-300.

1016. (SSF) Sodium:solute symporter family signatures and profile

PROSITE: PDOC00429. PROSITE cross-reference(s)PS00456; NA_SOLUT_SYMP_1

25 PS00457; NA_SOLUT_SYMP_2 PS50283; NA_SOLUTE_SYMP_3

It has been shown [1,2] that integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families is known as the sodium:solute symporter family (SSF) and

30 currently consists of the following proteins:

-Mammalian Na⁺/glucose co-transporter.

-Mammalian Na⁺/myo-inositol co-transporter.

-Mammalian Na⁺/nucleoside co-transporter.

831

- Mammalian Na⁺/neutral amino acid co-transporter.
- Escherichia coli Na⁺/proline symporter (gene putP).
- Escherichia coli Na⁺/pantothenate symporter (gene panF).
- Escherichia coli hypothetical protein yidK.
- 5 -Escherichia coli hypothetical protein yjcG.
- Bacillus subtilis hypothetical protein ywcA (ipa-31R).

These integral membrane proteins are predicted to comprise at least ten membrane spanning domains. Two conserved regions were selected as signature patterns; the first one is located in the fourth transmembrane region and the second one in a loop between two

10 transmembrane regions in the C-terminal part of these proteins.

Consensus pattern[GS]-x(2)-[LIY]-x(3)-[LIVMFYWSTAG][LIVMFYWSTAG SEQ ID NO:764](10)-[LIY]-[TAV]-x(2)-G-G-[LMF]-x-[SAP]. Sequences known to belong to this class detected by the patternALL.

15 Consensus pattern[GAST][GAST SEQ ID NO:179]-[LIVM][LIVM SEQ ID NO:4]-x(3)-[KR]-x(4)-G-A-x(2)-[GAS]-[LIVMGS][LIVMGS SEQ ID NO:765]-[LIVMW][LIVMW SEQ ID NO:235]-[LIVMGAT][LIVMGAT SEQ ID NO:766]-G-x-[LIVMGA][LIVMGA SEQ ID NO:175] Sequences known to belong to this class detected by the patternALL, except for E.coli yidK.

20 Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1]Reizer J., Reizer A., Saier M.H. Jr. Res. Microbiol. 141:1069-1072(1991).

25 [2]Reizer J., Reizer A., Saier M.H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).

1017. SurE - Survival protein SurE

E. coli cells with the surE gene disrupted are found to survive poorly in stationary phase [1].

It is suggested that SurE may be involved in stress response. Yeast also contains a member of

30 the family Swiss:P38254. Swiss:P30887 can complement a mutation in acid phosphatase, suggesting that members of this family could be phosphatases. Number of members: 17.

[1]Medline: 95014035. A new gene involved in stationary-phase survival located at 59 minutes on the Escherichia coli chromosome. Li C, Ichikawa JK, Ravetto JJ, Kuo HC, Fu JC, Clarke S; J Bacteriol 1994;176:6015-6022.

5 [2]Medline: 93046805. Complementation of Saccharomyces cerevisiae acid phosphatase mutation by a genomic sequence from the yeast Yarrowia lipolytica identifies a new phosphatase. Treton BY, Le Dall MT, Gaillardin CM; Curr Genet 1992;22:345-355.

1018. Synuclein - Synuclein

There are three types of synucleins in humans, these are called alpha, beta and gamma.

10 Alpha synuclein has been found mutated in families with autosomal dominant Parkinson's disease. A peptide of alpha synuclein has also been found in amyloid plaques in Alzheimer's patients. Number of members: 12.

[1]Medline: 98424410. The synuclein family. Lavedan C; Genome Res 1998;8:871-880.

15

1019. (T-box) T-box domain signatures

PROSITE: PDOC00972. PROSITE cross-reference(s) PS01283; TBOX_1 PS01264; TBOX_2

20 A number of eukaryotic DNA-binding proteins contain a domain of about 170 to 190 amino acids known as the T-box domain [1,2,3] and which probably binds DNA. The T-box has first been found in the mice T locus (Brachyury) protein, a transcription factor involved in mesoderm differentiation. It has since been found in the following proteins:

-Vertebrate and invertebrate homologs of the T protein.

-Mammalian proteins TBX1 to TBX6.

25 -Mammalian protein TBR1 which is expressed specifically in brain.

-Xenopus laevis eomesodermin (eomes).

-Xenopus laevis Vegt (or Antipodean), a transcription factor that activates the expression of wnt-8, eomes and Brachyury.

-Chicken TbxT.

30 -Drosophila protein optomotor-blind (omb).

-Drosophila protein brachyenteron (byn) (also known as Trg), which is required for the specification of the hindgut and anal pads.

-Drosophila protein H15.

-*Caenorhabditis elegans* protein tbx-12.

-*Caenorhabditis elegans* hypothetical proteins F21H11.3, F40H6.4, T07C4.2, T07C4.6 and ZK177.10.

5 Two conserved regions were selected as signature patterns for the T-domain. The first region corresponds to the N-terminal of the domain and the second one to the central part.

Consensus pattern L-W-x(2)-[FC]-x(3,4)-[NT]-E-M-[LIV](2)-T-x(2)-G-[RG]-[KRQ]

Sequences known to belong to this class detected by the pattern ALL, except for *C.elegans* ZK177.10.

10 Consensus pattern [LIVMYW][LIVMYW SEQ ID NO:767]-H-[PADH][PADH SEQ ID NO:768]-[DEN]-[GS]-x(3)-G-x(2)-W-M-x(3)-[IVA]-x- F Sequences known to belong to this class detected by the pattern ALL, except for *C.elegans* tbx-12, ZK177.10 and *Drosophila* H15.

15 [1] Bollag R.J., Siegfried Z., Cebra-Thomas J.A., Garvey N., Davison E.M., Silver L.M. Nat. Genet. 7:383-389(1994).

[2] Agulnik S.I., Garvey N., Hancock S., Ruvinsky I., Chapman D.L., Agulnik I., Bollag R.J., Papaioannou V.E., Silver L.M. Genetics 144:249-254(1996).

[3] Papaioannou V.E. Trends Genet. 13:212-213(1997).

20

1020. Toprim - Toprim domain

This is a conserved region from DNA primase. This corresponds to the Toprim domain common to DnaG primases, topoisomerases, OLD family nucleases and RecR proteins [1].

25 Both DnaG motifs IV and V are present in the alignment, the Dx(D) motif may be involved in Mg²⁺ binding and mutations to the conserved glutamate (IV) completely abolish DnaG type primase activity [1]. DNA primase EC:2.7.7.6 is a nucleotidyltransferase it synthesizes the oligoribonucleotide primers required for DNA replication on the lagging strand of the replication fork; it can also prime the leading strand and has been implicated in cell division [2]. Number of members: 133.

30

[1] Medline: 98391745. Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Aravind L, Leipe DD, Koonin EV; Nucleic Acids Res 1998;26:4205-4213.

[2]Medline: 97368180. Cloning and analysis of the dnaG gene encoding *Pseudomonas putida* DNA primase. Szafranski P, Smith CL, Cantor CR; *Biochim Biophys Acta* 1997;1352:243-248.

5 [3]Medline: 94124015. The *Haemophilus influenzae* dnaG sequence and conserved bacterial primase motifs. Versalovic J, Lupski JR; *Gene* 1993;136:281-286.

1021. TraB - TraB family

pAD1 is a hemolysin/bacteriocin plasmid originally identified in *Enterococcus faecalis* DS16. It encodes a mating response to a peptide sex pheromone, cAD1, secreted by recipient
10 bacteria. Once the plasmid pAD1 is acquired, production of the pheromone ceases--a trait related in part to a determinant designated traB. However a related protein is found in *C. elegans* Swiss:Q94217, suggesting that members of the TraB family have some more general function. Number of members: 12.

15 [1]Medline: 94302142. Characterization of the determinant (traB) encoding sex pheromone shutdown by the hemolysin/bacteriocin plasmid pAD1 in *Enterococcus faecalis*. An FY, Clewell DB; *Plasmid* 1994;31:215-221.

1022. (Transpo_mutator) Transposases, Mutator family, signature
20 PROSITE: PDOC00770. PROSITE cross-reference(s) PS01007;
TRANSPOSASE_MUTATOR

Autonomous mobile genetic elements such as transposon or insertion sequences (IS) encode an enzyme, called transposase, required for excising and inserting the mobile element. On the basis of sequence similarities, transposases can be grouped into various families. One
25 of these families has been shown [1,2,3,E1] to consist of transposases from the following elements:

- Mutator from Maize.
- Is1201 from *Lactobacillus helveticus*.
- Is905 from *Lactococcus lactis*.
- 30 -Is1081 from *Mycobacterium bovis*.
- Is6120 from *Mycobacterium smegmatis*.
- Is406 from *Pseudomonas cepacia*.
- IsRm3 from *Rhizobium meliloti*.

-IsRm5 from *Rhizobium meliloti*.

-Is256 from *Staphylococcus aureus*.

-IsT2 from *Thiobacillus ferrooxidans*.

5 The maize Mutator transposase (MudrA) is a protein of 823 amino acids; the bacterial transposases listed above are proteins of 300 to 420 amino acids. These proteins contain a conserved domain of about 130 residues; a signature pattern was derived from the most conserved part of this domain.

10 Consensus patternD-x(3)-G-[LIVMF][LIVMF SEQ ID NO:2]-x(6)-[STAV][STAV SEQ ID NO:105]-[LIVMFYW][LIVMFYW SEQ ID NO:26]-[PT]-x-[STAV][STAV SEQ ID NO:105]-x(2)-[QR]-x-C-x(2)-H. Sequences known to belong to this class detected by the patternALL.

[1]Eisen J.A., Benito M.-I., Walbot V. Nucleic Acids Res. 22:2634-2636(1994).

15 [2]Guilhot C., Gicquel B., Davies J., Martin C. Mol. Microbiol. 6:107-113(1992).

[3]Wood M.S., Byrne A., Lessie T.G. Gene 105:101-105(1991).

1023. Transposase_8 - Transposase

20 Transposase proteins are necessary for efficient DNA transposition. This family consists of various *E. coli* insertion elements and other bacterial transposases some of which are members of the IS3 family. Number of members: 58.

[1]Medline: 97324595. Genetic organization and transposition properties of IS511. D. A. Mullin, D. L. Zies, A. H. Mullin, N. Caballera & B. Ely; Mol Gen Genet 1997;254:456-463.

25 [2]Medline: 97128810. The use of an improved transposon mutagenesis system for DNA sequencing leads to the characterization of a new insertion sequence of *Streptomyces lividans* 66. J. Fischer, H. Maier, P. Viell & J. Altenbuchner; Gene 1996;180:81-89.

[3]Medline: 97074647. Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence ISRm6, a small transposable element that belongs to the IS3 family. S. Zekri & N. Toro; Gene 1996;175:43-48.

30

1024. tRNA_int_endo - tRNA intron endonuclease

Members of this family cleave pre tRNA at the 5' and 3' splice sites to release the intron
EC:3.1.27.9. Number of members: 8.

[1]Medline: 97344075. Properties of *H. volcanii* tRNA intron endonuclease reveal a
relationship between the archaeal and eucaryal tRNA intron processing systems. Kleman-
Leyer K, Armbruster DW, Daniels CJ; Cell 1997;89:839-847.

1025. Urease - Urease signatures

PROSITE: PDOC00133PROSITE cross-reference(s) PS01120; UREASE_1 PS00145;
UREASE_2

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea
to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in
1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease
is a hexamer of identical chains. In bacteria [2], it consists of either two or three different
subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a
carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the
catalytic mechanism [3].

As signatures for this enzyme, a region that contains two histidine that bind one of the
nickel ions and the region of the active site histidine was selected.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM][LIVM SEQ ID NO:4]-D-x-H-
[LIVM][LIVM SEQ ID NO:4]-H-x(3)-P [The two H's bind nickel].Sequences known to
belong to this class detected by the patternALL.

Consensus pattern[LIVM][LIVM SEQ ID NO:4](2)-[CT]-H-[HN]-L-x(3)-[LIVM][LIVM
SEQ ID NO:4]-x(2)-D-[LIVM][LIVM SEQ ID NO:4]-x-F-A [H is the active site residue].
Sequences known to belong to this class detected by the patternALL.

[1]Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

[2]Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3]Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

1026. Urease_beta - Urease beta subunit.

This subunit is known as alpha in *Helicobacter*. Number of members: 35.

[1] Medline: 95273988. The crystal structure of urease from *Klebsiella aerogenes*. Jabri E, Carr MB, Hausinger RP, Karplus PA; Science 1995;268:998-1004.

5

1027. UvrD-helicase - UvrD/REP helicase

The Rep family helicases are composed of four structural domains. The Rep family function as dimers. REP helicases catalyse ATP dependent unwinding of double stranded DNA to single stranded DNA. Swiss:P23478, Swiss:P08394 have large insertions near to the carboxy-terminus relative to other members of the family. Number of members: 52.

10

[1] Medline: 97433075. Major domain swiveling revealed by the crystal structures of complexes of *E. coli* Rep helicase bound to single-stranded DNA and ADP. Korolev S, Hsieh J, Gauss GH, Lohman TM, Waksman G; Cell 1997;90:635-647.

15

1028. V-type ATPase 116kDa subunit family (V_ATPase_sub_a)

This family consists of the 116kDa V-type ATPase (vacuolar (H⁺)-ATPases) subunits, as well as V-type ATP synthase subunit i. The V-type ATPases family are proton pumps that acidify intracellular compartments in eukaryotic cells for example yeast central vacuoles, clathrin-coated and synaptic vesicles. They have important roles in membrane trafficking processes [1]. The 116kDa subunit (subunit a) in the V-type ATPase is part of the V0 functional domain responsible for proton transport. The a subunit is a transmembrane glycoprotein with multiple putative transmembrane helices. It has a hydrophilic amino terminal and a hydrophobic carboxy terminal [1,2]. It has roles in proton transport and assembly of the V-type ATPase complex [1,2]. This subunit is encoded by two homologous genes in yeast VPH1 and STV1 [2].

20

25

Number of members: 27

[1] Forgac M; Medline: 99240666 "Structure and properties of the vacuolar (H⁺)-ATPases." J Biol Chem 1999;274:12951-12954.

[2] Forgac M; Medline: 99270697 "Structure and properties of the clathrin-coated vesicle and yeast vacuolar V-ATPases." J Bioenerg Biomembr 1999;31:57-65.

30

1029. Viral (Superfamily 1) RNA helicase (Viral_helicase1)

Number of members: 260

- 5 [1] Koonin EV, Dolja VV; Medline: 94094568 "Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences." Crit Rev Biochem Mol Biol 1993;28:375-430.

1030. Vesicular monoamine transporter (VMAT)

10

This family consists of various vesicular amine transporters with 12 transmembrane helices. These included vesicular acetylcholine transporters (VACHT) [3], and vesicular monoamine transporters (VMATs) [1,2] isoforms 1 adrenal and 2 brain (VMAT1 and VMAT2).

- 15 These proteins transport biogenic amines into synaptic vesicles or chromaffin granules [4]. VMATs pack monoamine neurotransmitters into secretory vesicles for regulated exocytotic release, they also protect against the parkinsonian neurotoxins MPP⁺ by transporting it into vesicles preventing it from acting on mitochondria [1].

- 20 Also in the family is *C. elegans* UNC-17 a putative vesicular acetylcholine transporter mutations in UNC-17 cause impaired neuromuscular function, giving rise to jerky or uncoordinated movement, [4].

Number of members: 15

- 25 [1] Krantz DE, Peter D, Liu Y, Edwards RH; Medline: 97197857 "Phosphorylation of a vesicular monoamine transporter by casein kinase II." J Biol Chem 1997;272:6752-6759.
[2] Erickson JD, Varoqui H, Schafer MK, Modi W, Diebler MF, Weihe E, Rand J, Eiden LE, Bonner TI, Usdin TB; Medline: 94350930 "Functional identification of a vesicular acetylcholine transporter and its expression from a 'cholinergic' gene locus." J Biol Chem
30 1994;269:21929-21932.
[3] Erickson JD, Schafer MK, Bonner TI, Eiden LE, Weihe E; Medline: 96209876 "Distinct pharmacological properties and distribution in neurons and endocrine cells of two isoforms of the human vesicular monoamine transporter." Proc Natl Acad Sci U S A 1996;93:5166-5171.

[4] Alfonso A, Grundahl K, Duerr JS, Han HP, Rand JB; Medline: 3342494 "The *Caenorhabditis elegans* unc-17 gene: a putative vesicular acetylcholine transporter." *Science* 1993;261:617-619.

- 5 1031. WW/rsp5/WWP domain signature and profile. Cross-reference(s): PS01159; WW_DOMAIN_1; PS50020; WW_DOMAIN_2

10 The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a
15 conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- 20 --Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin forms tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization
25 of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

--Utrophin, a dystrophin-like protein of unknown function.

- 30 --Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

--Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The

human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

--Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>), followed by a histidine-rich region, 3 WW domains and a HECT domain.

--Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

--Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

--Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

--IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

--Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each containing two WW-domains at the N-terminus.

--Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

--Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

Description of pattern(s) and/or profile(s):

Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE][GSTNE SEQ ID NO:737]-[GSTQCR][GSTQCR SEQ ID NO:738]-[FYW]-x(2)-P.

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

- [4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).
 [5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).
 [6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

5

1032. XPA protein signatures. cross-reference(s): XPA_1 PROSITE PS00752;
 PS00753;XPA_2.

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. XP-A is the most severe form of the disease and is due to defects in a 30 Kd nuclear protein called XPA (or XPAC) [2].

15

The sequence of the XPA protein is conserved from higher eukaryotes [3] to yeast (gene RAD14) [4]. XPA is a hydrophilic protein of 247 to 296 amino-acid residues which has a C4-type zinc finger motif in its central section.

20

Two signature were developed patterns for XPA proteins. The first corresponds to the zinc finger region, the second to a highly conserved region located some 12 residues after the zinc finger region.

25

Consensus pattern C-x-[DE]-C-x(3)-[LIVMF][LIVMF SEQ ID NO:2]-x(1,2)-D-x(2)-L-x(3)-
 F-x(4)-C-x(2)-C
 Consensus pattern [LIVM][LIVM SEQ ID NO:4](2)-T-[KR]-T-E-x-K-x-[DE]-Y-
 [LIVMF][LIVMF SEQ ID NO:2](2)-x-D-x-[DE]

30

- [1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).
 [2] Miura N., Miyamoto I., Asahina H., Satokata I., Tanaka K., Okada Y. J. Biol. Chem. 266:19786-19789(1991).
 [3] Shimamoto T., Kohno K., Tanaka K., Okada Y. Biochem. Biophys. Res. Commun. 181:1231-1237(1991).

[4] Bankmann M., Prakash L., Prakash S. Nature 355:555-558(1992).

1033. YCF9

5 This family consists of the hypothetical protein product of the YCF9 gene from chloroplasts and cyanobacteria. Number of members: 16

1034. (DUF15)

10 It is highly conserved between eubacteria and eukaryotes.

Number of members: 30

1035. Luminal portion of Cytochrome b559, alpha (gene psbE) subunit. (cytochr_b559a)

15 This family is the luminal portion of cytochrome b559 alpha chain, matches to this family should be accompanied by a match to the cytochr_b559 family also. The Prosite pattern matches the transmembrane region of the cytochrome b559 alpha and beta subunits. Number of members: 16

20

A. Asparaginase 2

25 Asparaginase II (L-asparagine aminohydrolase II) is an extracellular protein that may be associated with the cell wall and whose expression is affected by the availability of nitrogen. Asparaginase II catalyzes the reaction of L-Asparagine + H₂O = L-Aspartate + NH₃. As many leukemias have high requirements for aspartic acid, asparaginase II proteins are useful as reagents for screening compounds for activity as leukemia chemotherapy products. Asparaginase II protein can also be over- or under-expressed to alter amino acid content in
30 plant tissues or to modify nitrogen fixation and/or nitrogen metabolism in plants.

Ref: Bon et al. (1997) Appl Biochem Biotechnol 63-65: 203-12

B. Chloroa b-bind

Chlorophyll a-b binding proteins are located in the thylakoid membranes of the chloroplast and bind chlorophyll a and chlorophyll b, thereby triggering a chemical reaction

5 (photosynthesis). These proteins are useful in controlling the rate, efficiency and/or output of photosynthesis. Overexpression of chlorophyll a-b binding proteins is expected to increase the rate of photosynthesis.

Ref: Leutwiler et al. (1986) Nucleic Acids Res 14: 4051-64

10 Brandt et al. (1992) Plant Mol Biol 19: 699-703

C. DMRL synthase

DMRL Synthase (6,7-Dimethyl-8-Ribityllumazine Synthase) catalyzes the last step in
15 riboflavin (Vitamin B₂) synthesis, condensing 5-amino-6-(1'-D)-ribityl-amino-2,4(1H, 3H)-Pyrimidinedione with L-3,4-Dihydroxy-2-Butanone 4-Phosphate producing 6,7-Dimethyl-8-(1-D-Ribityl)Luminazine . The enzyme forms a homopentamer. Engineering of these proteins or those with homologous sequences/structures may allow control of the amounts of vitamin B₂ available in plants and/or accumulation of pigment, as well as altering reactions
20 requiring hydrogen ion carriers/transmitters.

Ref: Garcia-Ramirez et al. (1995) J Biol Chem 270: 23801-7

D. E1_N

25 These proteins are ATP-dependent DNA helicases that are required for initiation of viral DNA replication. They form a complex with the viral E2 protein. The E1-E2 complex binds to the replication origin that contains binding sites for both proteins. The majority of sequences known for this group of proteins are from various papillomaviruses, a type of
30 double stranded DNA virus. In plants, the prototype double stranded DNA virus is Cauliflower Mosaic virus (CaMV). Manipulation of these proteins, especially to produce variant proteins that form non-productive complexes, enables production of plants that are resistant to infection by double stranded DNA viruses.

Ref: Yang et al. (1993) PNAS USA **90**: 5086-90

Ustav and Stenlund (1991) EMBO J **10**: 449-57

Callaway et al. (1996) Mol Plant Microbe Interact **9**: 810-8

5

E. EF1_G

Elongation Factor-1 is composed of four subunits: alpha, beta, delta and gamma. Gamma subunits are presumed to play a role in anchoring the complex to other cellular components.

10

Studies of EF-1 genes in plants suggests that different forms of the EF-1 subunits may be expressed in particular organs or in response to stress. Manipulation of the activity of these proteins, either by altered expression level or by structural mutation, may result in the accumulation of a particular protein in a chosen organ or allow production of particular proteins during stress conditions.

15

Ref: Kinzy et al. (1994) NAR **22**: 2703-7

Dunn et al. (1993) Plant Mol Biol **23**: 221-5

Aguilar et al. (1991) Plant Mol Biol **17**: 351-60

20

F. ENV_polyprotein

This family comprises the envelope or coat proteins known from a number of different retroviruses. In mammalian species, retroviruses are responsible for diseases such as leukemia and HIV. In plants, retroviruses are known in both monocot (e.g. Zeon-1) and dicot (e.g. Arabidopsis and tobacco) species and have been shown to induce mutant alleles at new loci. Engineering of plant ENV proteins may allow mobilization or targeting of endogenous or introduced retroviruses, in essence generating a new method for mutant production, gene tagging and the like.

25

30

Ref: Mamoun et al (1990) J Virol **64**: 4180-8

Grandbastien et al. (1989) Nature **337**: 376-80

Wright and Voytas (1998) Genetics **149**: 703-15

G. Glycosyl_hydr9

Proteins having this domain (previously known as the glycosyl hydrolase family 5 domain) catalyze the endohydrolysis of 1,4- β -D-glucosidic linkages in cellulose. Numerous plant proteins with this domain exist and are expressed in an organ specific manner. They are involved in the fruit ripening process, in cell elongation and plant reproduction. Modulation of the activity of these proteins, either by over- or under-expression or by mutation of the polypeptide, could be used to affect post-harvest physiology (e.g. rate of ripening) or for engineering reproductive sterility.

Ref: Giorda et al. (1990) Biochemistry 29: 7264-9
Tucker et al. (1988) Plant Physiol 88: 1257-62
Shani et al. (1997) 43: 837-42
Milligan and Gasser (1995) Plant Mol Biol 28: 691-711

H. Glycosyl_hydr14

The β -amylases (family 14 of glycosyl hydrolases) catalyze the hydrolysis of 1,4- α -glucosidic linkages in polysaccharides and remove successive maltose units from the non-reducing ends of the chains. Mutants of β -amylase in Arabidopsis exhibited altered degradation of starch throughout the diurnal cycle. In addition, the mutant phenotypes indicated that these enzymes not only affect carbohydrate metabolism/catabolism, but also influence the amount of pigment stored within particular cells. Manipulation of the β -amylase genes enables control of plant pigmentation (for example, fibre pigment in cotton) as well as carbohydrate synthesis and degradation.

Ref: Zeeman et al. (1998) Plant J 15: 357-65
Hirano and Nakamura (1997) Plant Physiol 114: 5675-82
Kitamoto et al. (1988) J Bacteriol 170: 5848-54

I. Glycosyl_hydr15

Glycosyl hydrolases from family 15 (such as 1,4-Alpha-D-Glucan glucohydrolase,) catalyze the hydrolysis of terminal 1,4-linked alpha-D-glucose residues successively from the non-reducing ends of the chains resulting in the release of β -D-Glucose. In plants these proteins have been tied to the mobilization of the xyloglucan stored in the cotyledonary cell walls. Proteins such as these could be varied to affect the rate of plant growth (for example during germination), storage and/or use of glucose and other sugars by plant tissues and alteration of the properties, such as elasticity, of plant cell walls.

Ref: Crombie et al. (1998) Plant J 15: 27-38
Hata et al. (1991) Agric Biol Chem 55: 941-9

J. Glycosyl_hydr20

Members of the family 20 glycosyl hydrolases catalyze the hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl- β -D-hexosaminides. N-acetyl- β -glucosaminidase belongs to this family and exists in several different forms (consisting of various combinations of alpha and beta chains) depending on the organism. Family 20 glycosyl hydrolases have been implicated in lysosomal storage diseases (such as Sandhoff disease) and glycogen storage disease in humans. These types of proteins are also responsible for the hydrolysis of chitin. In plants, these proteins could be useful in controlling carbohydrate catabolism, thereby influencing the amount of sugars available for storage and/or use in other metabolic pathways. In addition, it is possible that such proteins could be used to engineer an endogenous insect protection mechanism, e.g. by secretion of a chitin-hydrolyzing composition by the plant.

Ref: Graham et al (1988) J Biol Chem 263: 16823-9
O'Dowd et al. (1988) Biochemistry 27: 5216-26

K. HMG box

The HMG box is a novel type of DNA-binding domain found in a diverse group of proteins. Numerous plant proteins contain this domain, such as the HMG1/2-like proteins. The expression of some of these HMG proteins appears to be regulated by circadian rhythms and in a light dependent manner, occurring at higher levels in roots, for example and lower levels in light-grown tissues such as cotyledons. Generally, HMG proteins are thought to influence transcription regulation. In plants, HMGs are believed to have a role in maintaining patterns of circadian-regulated expression for other genes, suggesting that these proteins could be exploited to control growth and development.

- 10 Ref: Laudet et al. (1993) Nucleic Acids Res 21: 2493-501
Zheng et al. (1993) Plant Mol Biol 23: 813-23
Grasser et al. (1993) Plant Mol Biol 23: 619-25

L. IL2

15

Interleukin-2 (IL-2) is produced in mammals by T cells in response to antigenic or mitogenic stimulation and is crucial for proper regulation and functioning of the immune response. IL-2 is capable of stimulating B cells, monocytes, lymphokine-activated killer cells, natural killer cells and glioma cells. Plant extracts have also been shown to stimulate the immune system (for example, mistletoe therapy for human cancer). It is known that IL-2 is involved in feedback inhibition pathways that impact the inflammatory response as well as the growth inhibition of tumor reactive T cells. Plant proteins containing IL-2-like sequences are useful as immunity-based therapeutics, acting in a manner similar to IL-2 in mammals.

- 20
25 Ref: Heike et al. (1997) Scand J Immunol 45: 221-6
Ariel et al. (1998) J Immunol 161: 2465-72
Schink (1997) Anticancer Drugs 8 Suppl 1: S47-51

M. Oxidored FMN

30

NADPH dehydrogenases catalyze the reaction $\text{NADPH} + \text{acceptor} = \text{NADP}(+) + \text{reduced acceptor}$. One member of this family is yeast "old yellow enzyme" (OYE) and is thought to be involved in oxylipin metabolism. A second yeast family member is a protein that binds

estrogen binding protein (EBP) in addition to exhibiting oxidoreductase activity. An Arabidopsis homolog to OYE has been described and estrogen binding proteins in plants have been reported. Plant proteins from this class have the potential to be used to modify lipid metabolism/catabolism. These proteins may also have use as therapeutics for breast and prostate cancer, and other abnormal growth in steroid-sensitive tissues.

Ref: Baker et al. (1998) Proc Soc Exp Biol Med 217: 317-21
Schaller and Weiler (1997) J Biol Chem 272: 28066-72
Mandani et al. (1994) PNAS USA 91: 922-6

N. Oxidored_q2

The NADH-plastoquinone oxidoreductases catalyze the reaction $\text{NADH} + \text{plastoquinone} = \text{NAD}(+) + \text{plastoquinol}$. In plants these reactions occur in the chloroplast and are believed to participate in a chloroplast respiratory system. Here, the NDH complex is postulated to act as a valve to remove excess reduction equivalents in the chloroplasts. Manipulation of these proteins may improve the rate or efficiency of photosynthesis.

Ref: Burrows et al. (1998) EMBO J 17: 868-76
Kofer et al (1998) Mol Gen Genet 258: 166-73
Maier et al. (1995) J Mol Biol 251: 614-28

O. PABP

Polyadenylate binding proteins bind the poly (A) tail of mRNA. Plants, as exemplified by Arabidopsis, contain numerous PABP genes that are expressed in an organ-specific manner. For example, PABP2 is functional in roots and shoots, while PABP5 is expressed predominantly in immature flowers. The PABP proteins are implicated in numerous aspects of posttranscriptional regulation including mRNA turnover and translational initiation. Control of activity of PABP proteins provides the ability to control the expression of various genes in particular organs during development.

Ref: Hilson et al (1993) Plant Physiol 103: 525-33

P. Parvo coat

5 Parvoviruses are linear single-stranded DNA viruses that are encapsulated by three capsid
proteins. Plants are susceptible to infection by single stranded DNA viruses such as Maize
streak virus (MSV) and various Gemini viruses. The coat proteins in these plant viruses are
critical to the virus life cycle within the plant. For example, the coat protein of MSV is
thought to be involved in intra- and inter-cellular movement within the plant. Engineering of
10 proteins having similarity to parvoviral coat proteins, especially to produce proteins that
interfere with maturation of the virus particle, enables the production of plants having better
resistance to natural plant single-stranded DNA viruses.

Ref: Liu et al. (1997) J Gen Virol 78: 1265-70

15 Rohde et al. (1990) Virology 176: 648-51

Q. Pkinase_C

Plant serine/threonine protein kinases possessing this domain are expressed in all tissues and
20 are known to undergo serine-specific autophosphorylation and specifically phosphorylate two
ribosomal proteins, P14 and P16. During development, these proteins predominate during
high metabolic activity in growing buds, root tips, leaf margins and germinating seeds. They
are thought to be involved in the control of plant growth and development. In addition, two
genes encoding proteins from this family have been described that help plant cells adapt
25 during cold or high salt stresses. Consequently, engineering Pkinase C proteins provides a
way to control general growth/development of the plant as well as a means to provide
endogenous protection against environmental stresses.

Ref: Zhang et al. (1994) J Biol Chem 269: 17586-92

30 Mizoguchi et al. (1995) FEBS Lett 358: 199-204

R. REV

The REV proteins act post-transcriptionally to relieve negative repression of GAG and ENV production in retroviruses such as Human Immunodeficiency Virus type I (HIV-1). Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutations at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant REV proteins enables control of transposition frequencies of corresponding transposable elements and provides a new tool for genetic engineering of plants.

Ref: Sodroski et al. (1986) Nature 321: 412-7
 Franchini et al. (1989) PNAS USA 86: 2433-7
 Marquet et al. (1995) 77: 113-24
 Grandbastien et al. (1989) Nature 337: 376-80
 Wright and Voytas (1998) Genetics 149: 703-15

S. RuBisCo small

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCo) catalyzes the initial step in the C3 photosynthetic carbon reduction cycle, adding carbon dioxide to D-ribulose 1,5-bisphosphate to form two molecules of 3-phospho-D-glycerate. RuBisCo is comprised of two subunits, one large which is synthesized in the chloroplast, and one small which is synthesized in the cytoplasm and then transported in to the chloroplast. The expression of the small subunit of RuBisCo is light regulated. Manipulation of these proteins could increase the efficiency of photosynthesis or allow alterations in developmental timing.

Ref: Giuliano et al. (1988) PNAS USA 85: 7089-93
 Dedonder et al. (1993) Plant Physiol 101: 801-8

T. Sialyltransf

Members of the CMP-N-acetylneuraminate- β -galactosamide- α -2,3-sialyltransferase family catalyze the following reaction:

CMP-N-acetylneuraminate + β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R = CMP + α -N-acetylneraminy-2,3- β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R. These proteins are thought to be responsible for the synthesis of the sequence neurac- α -2,3-gal- β -1,3-galnac- found on sugar chains)-linked to threonine or serine and also as a terminal
 5 sequence on certain gangliosides in mammalian cells. In plants, glycosyltransferases in the Golgi apparatus synthesize cell wall polysaccharides and elaborate the complex glycans of glycoproteins. Engineering of plant sialyltransferases allows targeting of proteins to particular cellular locations or enables the making of changes in cell wall structure.

- 10 Ref: Wee et al. (1998) Plant Cell 10: 1759-68
 Lee et al. (1994) J Biol Chem 269: 10028-33
 Kitagawa and Paulson (1994) J Biol Chem 269: 1394-401

U. Signal

15

Many plant proteins in this family contain sequences similar to those found in both components of the prokaryotic family of signal transducers known as the two-component systems. This suggests that activation may require a transfer of a phosphate group between the transmitter domain and the receiver domain. One family member in Arabidopsis appears
 20 to be involved in ethylene (a plant hormone) signal transduction. Other proteins in this family appear to be involved in the regulation of gene transcription under conditions of environmental stress. Signal proteins can be exploited to affect plant growth and development and/or control plant responses to stress conditions such as cold, nutrient availability, etc.

- 25 Ref: Chang et al. (1993) Science 262: 539-44
 Nagaya et al. (1993) Gene 131: 119-124
 Gottfert et al. (1990) PNAS USA 87: 2680-4

V. vMSA

30

vMSA proteins are major surface antigens presenting on the envelope of various retroviruses. Surface antigens of retroviruses are often involved in tropism of the virus. Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e.

transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutants at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant vMSA proteins enables control of tropism of plant retroviruses that might be used for genetic engineering tools, thus enabling targeting of the virus to particular species and/or tissues of plants.

Ref: Okamoto et al. (1988) J Gen Virol 69: 2575-83

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

W. zf-CCCH

This family of proteins is defined by having two CX(8)CX(5)CX(3)H-type zinc finger domains. These proteins cover a broad range of functions. For example, the COP1 protein acts as a repressor of photomorphogenesis in darkness; light stimuli abolish this suppressive action. In addition, COP1 protein can function as a negative transcriptional regulator capable of direct interaction with components of the G-protein signaling pathway. As a second example, a zf-CCCH protein identified in Arabidopsis appears to be involved in the resistance to DNA damage induced by UV light and chemical DNA-damaging agents.

Overexpression of this class of proteins permits production of plants that are better suited to adverse environments. Manipulation of expression of zf-CCCH proteins functioning as transcriptional regulators, such as COP1, enables manipulation of some signal transduction pathways.

Ref: Pang et al. (1993) Nucleic Acids Res 21: 1647-53

Deng et al. (1992) Cell 71: 791-801

X. zf-RanBP

Proteins falling within this category contain many X-X-F-G and X-F-X-F-G repeats, and may contain RANBP1-like or PPIase domains. Plant proteins having domains similar to these include PAS1 and GMSTI. PAS1 has been shown to have dramatic developmental affects that appear to be correlated with both cell division and cell wall elongation. GMSTI has high

identity to the yeast STI stress-inducible gene and has been shown to be heat inducible. Proteins such as these may be useful for controlling growth and form of development.

Ref: Vittorioso et al. (1998) Mol Cell Biol 18: 3034-43

5 Hernandez Torres et al. (1995) 27: 1221-6

Y. Peptidase M48.

10 Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are located in the membranes of the endoplasmic reticulum. They function in NH₂-terminal proteolytic processing, as shown for the yeast STE24 gene product. This gene is required for the correct processing of α -factor, a yeast pheromone. Family M48 peptidases also appear to be required for some prenylation reactions, mediating COOH-terminal CAAX processing. Prenylation reactions are believed to be involved in the regulation of protein-

15 protein and protein-membrane interactions. As an example, RAS GTPase activity is regulated in part by localization to the inner side of the plasma membrane upon prenylation. In plants, proteins from this family could be involved in pollen-stigma interactions such as those mediating self-pollination vs. outcrossing, or could be members of several secondary metabolism pathways.

20

Ref: Fujimura-Kamada et al. (1997) J Cell Biol. 136: 271-85. Tam et al. (1998) J Cell Biol. 142: 635-49.

Z. DNA Pol Viral N

25 The DNA pol Viral N domain is located at the N-terminal region of DNA polymerase isolated from several retrovirus viruses such as the Cauliflower Mosaic Virus. The domain motif has also been found in numerous other species from humans to cyanobacteria. In these organisms, this motif seems to be associated with two types of sequences; retrotransposons and mitochondrial genes. In the mitochondrial sequences this domain is potentially involved

30 in the self-splicing conducted by group II introns. Various manipulations of this gene in plants allows control of the numerous retrotransposons endogenous to plant genomes or allows engineering of mitochondrial function, especially to increase efficiency of energy utilization by cells.

REF: Chapdelaine and Bonen (1991) Cell 65: 465-72

Ferat and Miche (1993) Nature 364: 358-61

Wilson et al. (1994) 368: 32-8

5 Cambareri et al. (1994) 242: 658-65

Gaardner et al. (1981) NAR 9: 2871-2888

Cummings et al. (1990) Curr Genet 17: 375-402

Hattori et al. (1986) Nature 321: 625-8

10 Aa. Calpain_inhib

This domain is found in calpastatin, an inhibitor protein specific for calpain. Calpain is a non-lysosomal calcium-dependent intracellular protease that appears to be involved in the dynamic changes of the cytoskeleton, especially actin-related structures, during early *Drosophila* embryogenesis [1]. Calpastatins co-exist in cells with calpains and the subcellular
15 distribution of calpastatin is thought to be important to calpain regulation [2]. In plants calpains and calpastatins could be involved in embryogenesis and non-embryogenic organ reiteration. Mutations occurring in calpain inhibitor repeat domains would produce developmental abnormalities such as abnormal leaf, root or flower development.

20 Refs

1 Emori Y and Saigo K (1994) J Biol Chem 269: 25137-42.

2 Mellgren RL, Lane RD, Mericle MT (1989) Biochim Biophys Acta 999: 71-77.

Ab. chorismate_bind

25 Chorismate binding domains are present in plant anthranilate synthase (AS) genes. AS genes catalyze the first step in the biosynthesis of tryptophan by converting chorismate and L-glutamine to anthranilate, pyruvate and L-glutamate. Some of these genes are involved in feedback inhibition by tryptophan [1] while some are feedback insensitive [2]. In Arabidopsis, two AS genes have overlapping, but different distributions. One of these AS
30 genes is induced by wounding and bacterial pathogen infiltration [1]. Mutations in the chorismate binding domain would affect the production of tryptophan and could influence the plant's defense system. AS gene products can be used for *in vitro* synthesis of tryptophan and tryptophan derivatives.

Refs

- 1 Niyogi KK, Fink GR (1992) Plant Cell 4: 721-33.
2 Song HS, Brotherton JE, Gonzales RA, Wilholm JM (1998) Plant Physiol 117:533-
5 43.

Ac. late protein L2

Papillomaviruses are encapsulated double stranded DNA viruses. Plants are susceptible to infection by double stranded DNA viruses such as Cauliflower Mosaic virus (CaMV). The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of CaMV is thought to be involved in intra- and inter-cellular movement within the plant [1]. Engineering of proteins having similarity to papillomavirus coat proteins may enable the production of plants having better resistance to natural plant double stranded DNA viruses.

Refs

- 1 Thompson SR, Melcher U (1993) J Gen Virol 74: 1141-8.

Ad. Peptidase M41

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are integral membrane proteins. They seem to be involved in the degradation of carboxy-terminal-tagged cytoplasmic proteins. In plants, these proteins are located in the thylakoid membranes of the chloroplasts, their expression is light regulated and they are thought to be involved in degradation of soluble stromal proteins and turn-over of thylakoid proteins [1]. Manipulation of expression and structure of these proteins would have effects on the efficiency of photosynthesis and the development of chloroplasts.

Refs

- 1 Lindahl M, Tabak s, Cseke L, Pichersky E, Andersson B, Adam Z (1996) J Biol
30 Chem 271: 29329-34.

Ae. UPF0051

There is some evidence that, in plants, proteins in this family are involved in ATP synthesis in chloroplasts [1, 2]. Mutations in these proteins or altering their expression would affect the efficiency of photosynthesis and energy production.

5 Refs

- 1 Kostrzewa M, Zetsche K (1992) J Mol Biol 227: 961-70.
- 2 Kostrzewa M, Zetsche K (1993) Plant Mol Biol 23: 67-76

Af. E7

- 10 Papillomaviruses are encapsulated double stranded DNA viruses. The Papillomavirus early protein 7 (E7) is known as a potent immortalizing and transforming agent. Transformation by E7 is thought to be mediated by the physical association of E7 with cellular proteins regulating entry into the cell cycle [1]. The result is entry into the cell cycle and suppression of terminal differentiation in mammalian cells. Thus, engineering of proteins having
- 15 similarity to papillomavirus E7 protein enables the production of plants having altered cellular proliferation characteristics and possibly altered morphology. For example, overexpression of E7-like proteins would be expected to result in proliferation of cells of the tissue in which the E7 protein is expressed, perhaps with suppression of differentiation events. Thus, for example, overexpression of E7-like proteins in meristem cells can result in
- 20 taller plants and suppression of leafing and/or flowering.

Refs

- 1 Zwerschke W, Jansen-Durr P Adv Cancer Res 2000;78:1-29

25 Ag. Peptidase U7

- This protein is known to be an integral membrane protein in the cyanobacterium *Synechocystis* where it functions to digest cleaved signal peptides [1]. This activity is necessary to maintain proper secretion of mature proteins across the membrane. In higher plants this protein may be present in the plastid or chloroplast membranes where it would
- 30 function by enabling protein movement into and out of the chloroplasts. Mutations in this protein would be expected to affect the development of plastids, including chloroplasts, or alter the energy transfer system within the chloroplasts, thereby affecting growth and development.

Refs

- 1 Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N,
 Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A,
 Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A,
 5 Yamada M, Yasuda M, Tabata S (1996) DNA Res 3:109-36.

Ah. 5'-3' Exonuclease

The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA
 repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and
 10 also includes a domain that interacts with the mismatch repair protein MSH2. The 5'-3'
 exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast
 RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in
 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA
 repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence
 15 (SEQ ID NO: 769):

IMKKKLLLVDGSSLAFFALPPLTNSAGEPTNAVYGFLLKMLIKLIEQECPHIAVV
 FDAKAKTFRHELYEGYKAGRAP
 TPDELREQUIPLIKELLDALGIPLLEVAGYEADDVIGTLAKLAEKEGYEVLIVTGDRDLL
 20 QLVSDHVTVIITKKGIAEFTL
 FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTAALKLLQEYGSLEGIYANL
 DKLKGKKLREKLLAHKEDAKL
 SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE

25 Ref:

Fiorentini P. et al. RT. Mol. Cell. Biol. 17:2764-2773(1997).
 Tishkoff et al. Cancer Res. 0:0-0(1998).
 Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993).

AA. Activities of Polypeptides Comprising Signal Peptides

Polypeptides comprising signal peptides are a family of proteins that are typically targeted to (1) a particular organelle or intracellular compartment, (2) interact with a particular molecule or (3) for secretion outside of a host cell. Example of polypeptides comprising signal peptides include, without limitation, secreted proteins, soluble proteins, receptors, proteins retained in the ER, etc.

These proteins comprising signal peptides are useful to modulate ligand-receptor interactions, cell-to-cell communication, signal transduction, intracellular communication, and activities and/or chemical cascades that take part in an organism outside or within of any particular cell.

One class of such proteins are soluble proteins which are transported out of the cell. These proteins can act as ligands that bind to receptor to trigger signal transduction or to permit communication between cells.

Another class is receptor proteins which also comprise a retention domain that lodges the receptor protein in the membrane when the cell transports the receptor to the surface of the cell. Like the soluble ligands, receptors can also modulate signal transduction and communication between cells.

In addition the signal peptide itself can serve as a ligand for some receptors. An example is the interaction of the ER targeting signal peptide with the signal recognition particle (SRP). Here, the SRP binds to the signal peptide, halting translation, and the resulting SRP complex then binds to docking proteins located on the surface of the ER, prompting transfer of the protein into the ER.

A description of signal peptide residue composition is described below in Subsection IV.C.1.

III. Methods of Modulating Polypeptide Production

It is contemplated that polynucleotides of the invention can be incorporated into a host cell or in-vitro system to modulate polypeptide production. For instance, the SDFs prepared as described herein can be used to prepare expression cassettes useful in a number of techniques for suppressing or enhancing expression.

An example are polynucleotides comprising sequences to be transcribed, such as coding sequences, of the present invention can be inserted into nucleic acid constructs to modulate polypeptide production. Typically, such sequences to be transcribed are heterologous to at least one element of the nucleic acid construct to generate a chimeric gene or construct.

Another example of useful polynucleotides are nucleic acid molecules comprising regulatory sequences of the present invention. Chimeric genes or constructs can be generated when the regulatory sequences of the invention linked to heterologous sequences in a vector construct. Within the scope of invention are such chimeric gene and/or constructs.

Also within the scope of the invention are nucleic acid molecules, whereof at least a part or fragment of these DNA molecules are presented in TABLE 1 of the present application, and wherein the coding sequence is under the control of its own promoter and/or its own regulatory elements. Such molecules are useful for transforming the genome of a host cell or an organism regenerated from said host cell for modulating polypeptide production.

Additionally, a vector capable of producing the oligonucleotide can be inserted into the host cell to deliver the oligonucleotide.

More detailed description of components to be included in vector constructs are described both above and below.

Whether the chimeric vectors or native nucleic acids are utilized, such polynucleotides can be incorporated into a host cell to modulate polypeptide production. Native genes and/or nucleic acid molecules can be effective when exogenous to the host cell.

Methods of modulating polypeptide expression includes, without limitation:

Suppression methods, such as

Antisense

Ribozymes

Co-suppression

Insertion of Sequences into the Gene to be Modulated

Regulatory Sequence Modulation.

as well as Methods for Enhancing Production, such as
Insertion of Exogenous Sequences; and
5 Regulatory Sequence Modulation.

III.A. Suppression

Expression cassettes of the invention can be used to suppress expression of
endogenous genes which comprise the SDF sequence. Inhibiting expression can be useful,
for instance, to tailor the ripening characteristics of a fruit (Oeller et al., *Science* 254:437
10 (1991)) or to influence seed size_(WO98/07842) or to provoke cell ablation (Mariani et al.,
Nature 357: 384-387 (1992)).

As described above, a number of methods can be used to inhibit gene expression in
plants, such as antisense, ribozyme, introduction of exogenous genes into a host cell,
insertion of a polynucleotide sequence into the coding sequence and/or the promoter of the
15 endogenous gene of interest, and the like.

III.A.1. Antisense

An expression cassette as described above can be transformed into host cell or
plant to produce an antisense strand of RNA. For plant cells, antisense RNA inhibits gene
expression by preventing the accumulation of mRNA which encodes the enzyme of interest, *see*,
20 e.g., Sheehy et al., *Proc. Nat. Acad. Sci. USA*, 85:8805 (1988), and Hiatt et al., U.S. Patent No.
4,801,340.

III.A.2. Ribozymes

Similarly, ribozyme constructs can be transformed into a plant to cleave mRNA
and down-regulate translation.

25 III.A.3. Co-Suppression

Another method of suppression is by introducing an exogenous copy of the gene
to be suppressed. Introduction of expression cassettes in which a nucleic acid is configured in
the sense orientation with respect to the promoter has been shown to prevent the accumulation of
mRNA. A detailed description of this method is described above.

III.A.4. Insertion of Sequences into the Gene to be Modulated

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Homologous recombination could be used to target a polynucleotide insert to a gene using the Cre-Lox system (A.C. Vergunst et al., *Nucleic Acids Res.* 26:2729 (1998), A.C. Vergunst et al., *Plant Mol. Biol.* 38:393 (1998), H. Albert et al., *Plant J.* 7:649 (1995)).

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred for identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or any transgenic plants having a desired phenotype.

III.A.5. Regulatory Sequence Modulation

The SDFs described in Table 1, and fragments thereof are examples of nucleotides of the invention that contain regulatory sequences that can be used to suppress or inactivate transcription and/or translation from a gene of interest as discussed in I.C.5.

III.A.6. Genes Comprising Dominant-Negative Mutations

When suppression of production of the endogenous, native protein is desired it is often helpful to express a gene comprising a dominant negative mutation. Production of protein variants produced from genes comprising dominant negative mutations is a useful tool for research. Genes comprising dominant negative mutations can produce a variant polypeptide which is capable of competing with the native polypeptide, but which does not produce the native result. Consequently, over expression of genes comprising these mutations can titrate out an undesired activity of the native protein. For example, The product from a gene comprising a dominant negative mutation of a receptor can be used to constitutively activate or suppress a signal transduction cascade, allowing examination of the phenotype and thus the trait(s) controlled by that receptor and pathway. Alternatively, the protein arising from the gene comprising a dominant-negative mutation can be an inactive enzyme still capable

of binding to the same substrate as the native protein and therefore competes with such native protein.

Products from genes comprising dominant-negative mutations can also act upon the native protein itself to prevent activity. For example, the native protein may be active only as a homo-multimer or as one subunit of a hetero-multimer. Incorporation of an inactive subunit into the multimer with native subunit(s) can inhibit activity.

Thus, gene function can be modulated in host cells of interest by insertion into these cells vector constructs comprising a gene comprising a dominant-negative mutation.

III.B. Enhanced Expression

Enhanced expression of a gene of interest in a host cell can be accomplished by either (1) insertion of an exogenous gene; or (2) promoter modulation.

III.B.1. Insertion of an Exogenous Gene

Insertion of an expression construct encoding an exogenous gene can boost the number of gene copies expressed in a host cell.

Such expression constructs can comprise genes that either encode the native protein that is of interest or that encode a variant that exhibits enhanced activity as compared to the native protein. Such genes encoding proteins of interest can be constructed from the sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto.

Such an exogenous gene can include either a constitutive promoter permitting expression in any cell in a host organism or a promoter that directs transcription only in particular cells or times during a host cell life cycle or in response to environmental stimuli.

III.B.2. Regulatory Sequence Modulation

The SDFs of Table 1, and fragments thereof, contain regulatory sequences that can be used to enhance expression of a gene of interest. For example, some of these sequences contain useful enhancer elements. In some cases, duplication of enhancer elements or insertion of exogenous enhancer elements will increase expression of a desired gene from a particular promoter. As other examples, all II promoters require binding of a regulatory protein to be activated, while some promoters may need a protein that signals a promoter binding protein to expose a polymerase binding site. In either case, over-production of such proteins can be used to enhance expression of a gene of interest by increasing the activation time of the promoter.

Such regulatory proteins are encoded by some of the sequences in TABLE 1, fragments thereof, and substantially similar sequences thereto.

Coding sequences for these proteins can be constructed as described above.

5 IV. Gene Constructs and Vector Construction

To use isolated SDFs of the present invention or a combination of them or parts and/or mutants and/or fusions of said SDFs in the above techniques, recombinant DNA vectors which comprise said SDFs and are suitable for transformation of cells, such as plant cells, are usually prepared. The SDF construct can be made using standard recombinant DNA techniques
10 (Sambrook et al. 1989) and can be introduced to the species of interest by *Agrobacterium*-mediated transformation or by other means of transformation (e.g., particle gun bombardment) as referenced below.

The vector backbone can be any of those typical in the art such as plasmids, viruses, artificial chromosomes, BACs, YACs and PACs and vectors of the sort described by

- 15 (a) **BAC:** Shizuya et al., Proc. Natl. Acad. Sci. USA 89: 8794-8797 (1992); Hamilton et al., Proc. Natl. Acad. Sci. USA 93: 9975-9979 (1996);
- (b) **YAC:** Burke et al., Science 236:806-812 (1987);.
- (c) **PAC:** Sternberg N. et al., Proc Natl Acad Sci U S A. Jan;87(1):103-7 (1990);
- (d) **Bacteria-Yeast Shuttle Vectors:** Bradshaw et al., Nucl Acids Res 23: 4850-
20 4856 (1995);
- (e) **Lambda Phage Vectors:** Replacement Vector, e.g., Frischauf et al., J. Mol Biol 170: 827-842 (1983); or Insertion vector, e.g., Huynh et al., In: Glover NM (ed) DNA Cloning: A practical Approach, Vol.1 Oxford: IRL Press (1985);
- 25 (f) **T-DNA gene fusion vectors :**Walden et al., Mol Cell Biol 1: 175-194 (1990); and
- (g) **Plasmid vectors:** Sambrook et al., infra.

Typically, a vector will comprise the exogenous gene, which in its turn comprises an SDF of the present invention to be introduced into the genome of a host cell, and which gene
30 may be an antisense construct, a ribozyme construct chimera, or a coding sequence with any desired transcriptional and/or translational regulatory sequences, such as promoters, UTRs,

and 3' end termination sequences. Vectors of the invention can also include origins of replication, scaffold attachment regions (SARs), markers, homologous sequences, introns, etc.

A DNA sequence coding for the desired polypeptide, for example a cDNA sequence encoding a full length protein, will preferably be combined with transcriptional and translational initiation regulatory sequences which will direct the transcription of the sequence from the gene in the intended tissues of the transformed plant.

For example, for over-expression, a plant promoter fragment may be employed that will direct transcription of the gene in all tissues of a regenerated plant. Alternatively, the plant promoter may direct transcription of an SDF of the invention in a specific tissue (tissue-specific promoters) or may be otherwise under more precise environmental control (inducible promoters).

If proper polypeptide production is desired, a polyadenylation region at the 3'-end of the coding region is typically included. The polyadenylation region can be derived from the natural gene, from a variety of other plant genes, or from T-DNA.

The vector comprising the sequences from genes or SDF or the invention may comprise a marker gene that confers a selectable phenotype on plant cells. The vector can include promoter and coding sequence, for instance. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to kanamycin, G418, bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or phosphinotricin.

IV.A. Coding Sequences

Generally, the sequence in the transformation vector and to be introduced into the genome of the host cell does not need to be absolutely identical to an SDF of the present invention. Also, it is not necessary for it to be full length, relative to either the primary transcription product or fully processed mRNA. Furthermore, the introduced sequence need not have the same intron or exon pattern as a native gene. Also, heterologous non-coding segments can be incorporated into the coding sequence without changing the desired amino acid sequence of the polypeptide to be produced.

IV.B. Promoters

As explained above, introducing an exogenous SDF from the same species or an orthologous SDF from another species can modulate the expression of a native gene

corresponding to that SDF of interest. Such an SDF construct can be under the control of either a constitutive promoter or a highly regulated inducible promoter (e.g., a copper inducible promoter). The promoter of interest can initially be either endogenous or heterologous to the species in question. When re-introduced into the genome of said species, such promoter becomes exogenous to said species. Over-expression of an SDF transgene can lead to co-suppression of the homologous endogeneous sequence thereby creating some alterations in the phenotypes of the transformed species as demonstrated by similar analysis of the chalcone synthase gene (Napoli et al., *Plant Cell* 2:279 (1990) and van der Krol et al., *Plant Cell* 2:291 (1990)). If an SDF is found to encode a protein with desirable characteristics, its over-production can be controlled so that its accumulation can be manipulated in an organ- or tissue-specific manner utilizing a promoter having such specificity.

Likewise, if the promoter of an SDF (or an SDF that includes a promoter) is found to be tissue-specific or developmentally regulated, such a promoter can be utilized to drive or facilitate the transcription of a specific gene of interest (e.g., seed storage protein or root-specific protein). Thus, the level of accumulation of a particular protein can be manipulated or its spatial localization in an organ- or tissue- specific manner can be altered.

IV. C Signal Peptides

SDFs of the present invention containing signal peptides are indicated in Table 1. In some cases it may be desirable for the protein encoded by an introduced exogenous or orthologous SDF to be targeted (1) to a particular organelle intracellular compartment, (2) to interact with a particular molecule such as a membrane molecule or (3) for secretion outside of the cell harboring the introduced SDF. This will be accomplished using a signal peptide.

Signal peptides direct protein targeting, are involved in ligand-receptor interactions and act in cell to cell communication. Many proteins, especially soluble proteins, contain a signal peptide that targets the protein to one of several different intracellular compartments. In plants, these compartments include, but are not limited to, the endoplasmic reticulum (ER), mitochondria, plastids (such as chloroplasts), the vacuole, the Golgi apparatus, protein storage vesicles (PSV) and, in general, membranes. Some signal peptide sequences are conserved, such as the Asn-Pro-Ile-Arg amino acid motif found in the N-terminal propeptide signal that targets proteins to the vacuole (Marty (1999) *The Plant Cell* 11: 587-599). Other signal peptides do not have a consensus sequence *per se*, but are largely composed of

hydrophobic amino acids, such as those signal peptides targeting proteins to the ER (Vitale and Denecke (1999) *The Plant Cell* 11: 615-628). Still others do not appear to contain either a consensus sequence or an identified common secondary sequence, for instance the chloroplast stromal targeting signal peptides (Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). Furthermore, some targeting peptides are bipartite, directing proteins first to an organelle and then to a membrane within the organelle (e.g. within the thylakoid lumen of the chloroplast; see Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). In addition to the diversity in sequence and secondary structure, placement of the signal peptide is also varied. Proteins destined for the vacuole, for example, have targeting signal peptides found at the N-terminus, at the C-terminus and at a surface location in mature, folded proteins. Signal peptides also serve as ligands for some receptors.

These characteristics of signal proteins can be used to more tightly control the phenotypic expression of introduced SDFs. In particular, associating the appropriate signal sequence with a specific SDF can allow sequestering of the protein in specific organelles (plastids, as an example), secretion outside of the cell, targeting interaction with particular receptors, etc. Hence, the inclusion of signal proteins in constructs involving the SDFs of the invention increases the range of manipulation of SDF phenotypic expression. The nucleotide sequence of the signal peptide can be isolated from characterized genes using common molecular biological techniques or can be synthesized in vitro.

In addition, the native signal peptide sequences, both amino acid and nucleotide, described in Table 1 can be used to modulate polypeptide transport. Further variants of the native signal peptides described in Table 1 are contemplated. Insertions, deletions, or substitutions can be made. Such variants will retain at least one of the functions of the native signal peptide as well as exhibiting some degree of sequence identity to the native sequence.

Also, fragments of the signal peptides of the invention are useful and can be fused with other signal peptides of interest to modulate transport of a polypeptide.

V. Transformation Techniques

A wide range of techniques for inserting exogenous polynucleotides are known for a number of host cells, including, without limitation, bacterial, yeast, mammalian, insect and plant cells.

Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature. See, e.g. Weising et al., *Ann. Rev. Genet.* 22:421 (1988); and Christou, *Euphytica*, v. 85, n.1-3:13-27, (1995).

DNA constructs of the invention may be introduced into the genome of the desired plant host by a variety of conventional techniques. For example, the DNA construct may be introduced directly into the genomic DNA of the plant cell using techniques such as electroporation and microinjection of plant cell protoplasts, or the DNA constructs can be introduced directly to plant tissue using ballistic methods, such as DNA particle bombardment. Alternatively, the DNA constructs may be combined with suitable T-DNA flanking regions and introduced into a conventional *Agrobacterium tumefaciens* host vector. The virulence functions of the *Agrobacterium tumefaciens* host will direct the insertion of the construct and adjacent marker into the plant cell DNA when the cell is infected by the bacteria (McCormac et al., *Mol. Biotechnol.* 8:199 (1997); Hamilton, *Gene* 200:107 (1997)); Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983).

Microinjection techniques are known in the art and well described in the scientific and patent literature. The introduction of DNA constructs using polyethylene glycol precipitation is described in Paszkowski et al. *EMBO J.* 3:2717 (1984). Electroporation techniques are described in Fromm et al. *Proc. Natl Acad. Sci. USA* 82:5824 (1985). Ballistic transformation techniques are described in Klein et al. *Nature* 327:773 (1987). *Agrobacterium tumefaciens*-mediated transformation techniques, including disarming and use of binary or co-integrate vectors, are well described in the scientific literature. See, for example Hamilton, *CM., Gene* 200:107 (1997); Müller et al. *Mol. Gen. Genet.* 207:171 (1987); Komari et al. *Plant J.* 10:165 (1996); Venkateswarlu et al. *Biotechnology* 9:1103 (1991) and Gleave, *AP., Plant Mol. Biol.* 20:1203 (1992); Graves and Goldman, *Plant Mol. Biol.* 7:34 (1986) and Gould et al., *Plant Physiology* 95:426 (1991).

Transformed plant cells which are derived by any of the above transformation techniques can be cultured to regenerate a whole plant that possesses the transformed genotype and thus the desired phenotype such as seedlessness. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., *Protoplasts Isolation and Culture* in "Handbook of Plant Cell Culture," pp. 124-176, MacMillan Publishing Company, New York, 1983; and Binding, *Regeneration of Plants, Plant Protoplasts*, pp. 21-73,

CRC Press, Boca Raton, 1988. Regeneration can also be obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. *Ann. Rev. of Plant Phys.* **38**:467 (1987). Regeneration of monocots (rice) is described by Hosoyama et al. (*Biosci. Biotechnol. Biochem.* **58**:1500 (1994)) and by Ghosh et al. (*J. Biotechnol.* **32**:1 (1994)). The nucleic acids of the invention can be used to confer desired traits on essentially any plant.

Thus, the invention has use over a broad range of plants, including species from the genera *Anacardium*, *Arachis*, *Asparagus*, *Atropa*, *Avena*, *Brassica*, *Citrus*, *Citrullus*, *Capsicum*, *Carthamus*, *Cocos*, *Coffea*, *Cucumis*, *Cucurbita*, *Daucus*, *Elaeis*, *Fragaria*, *Glycine*, *Gossypium*, *Helianthus*, *Heterocallis*, *Hordeum*, *Hyoscyamus*, *Lactuca*, *Linum*, *Lolium*, *Lupinus*, *Lycopersicon*, *Malus*, *Manihot*, *Majorana*, *Medicago*, *Nicotiana*, *Olea*, *Oryza*, *Panicum*, *Pannisetum*, *Persea*, *Phaseolus*, *Pistachia*, *Pisum*, *Pyrus*, *Prunus*, *Raphanus*, *Ricinus*, *Secale*, *Senecio*, *Sinapis*, *Solanum*, *Sorghum*, *Theobromus*, *Trigonella*, *Triticum*, *Vicia*, *Vitis*, *Vigna*, and, *Zea*.

One of skill will recognize that after the expression cassette is stably incorporated in transgenic plants and confirmed to be operable, it can be introduced into other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

The particular sequences of SDFs identified are provided in the attached TABLE 1. One of ordinary skill in the art, having this data, can obtain cloned DNA fragments, synthetic DNA fragments or polypeptides constituting desired sequences by recombinant methodology known in the art or described herein.

EXAMPLES

The invention is illustrated by way of the following examples. The invention is not limited by these examples as the scope of the invention is defined solely by the claims following.

EXAMPLE 1: cDNA PREPARATION

A number of the nucleotide sequences disclosed in TABLE 1 herein as representative of the SDFs of the invention can be obtained by sequencing genomic DNA (gDNA) and/or cDNA from corn plants grown from HYBRID SEED # 35A19, purchased from Pioneer Hi-Bred International, Inc., Supply Management, P.O. Box 256, Johnston, Iowa 50131-0256.

A number of the nucleotide sequences disclosed in TABLE 1 herein as representative of the SDFs of the invention can also be obtained by sequencing genomic DNA from *Arabidopsis thaliana*, Wassilewskija ecotype or by sequencing cDNA obtained from mRNA from such plants as described below. This is a true breeding strain. Seeds of the plant are
5 available from the Arabidopsis Biological Resource Center at the Ohio State University, under the accession number CS2360. Seeds of this plant were deposited under the terms and conditions of the Budapest Treaty at the American Type Culture Collection, Manassas, VA on August 31, 1999, and were assigned ATCC No. PTA-595.

Other methods for cloning full-length cDNA are described, for example, by Seki et al., *Plant Journal* 15:707-720 (1998) "High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated Cap trapper"; Maruyama et al., *Gene* 138:171 (1994) "Oligo-capping a
10 simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides"; and WO 96/34981.

Tissues were, or each organ was, individually pulverized and frozen in liquid
15 nitrogen. Next, the samples were homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed. Then the sample was applied to a 2M sucrose cushion to isolate polysomes. The RNA was isolated by treatment with detergents and proteinase K followed by ethanol precipitation and
20 centrifugation. The polysomal RNA from the different tissues was pooled according to the following mass ratios: 15/15/1 for male inflorescences, female inflorescences and root, respectively. The pooled material was then used for cDNA synthesis by the methods described below.

Starting material for cDNA synthesis for the exemplary corn cDNA clones
25 with sequences presented in TABLE 1 was poly(A)-containing polysomal mRNAs from inflorescences and root tissues of corn plants grown from HYBRID SEED # 35A19. Male inflorescences and female (pre-and post-fertilization) inflorescences were isolated at various stages of development. Selection for poly(A) containing polysomal RNA was done using oligo d(T) cellulose columns, as described by Cox and Goldberg, "Plant Molecular Biology:
30 A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The quality and the integrity of the polyA+ RNAs were evaluated.

Starting material for cDNA synthesis for the exemplary *Arabidopsis* cDNA clones with sequences presented in TABLE 1 was polysomal RNA isolated from the top-most inflorescence tissues of *Arabidopsis thaliana* Wassilewskija (Ws.) and from roots of *Arabidopsis thaliana* Landsberg erecta (L. er.), also obtained from the Arabidopsis

5 Biological Resource Center. Nine parts inflorescence to every part root was used, as measured by wet mass. Tissue was pulverized and exposed to liquid nitrogen. Next, the sample was homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed and the sample was applied to a 2M
10 sucrose cushion to isolate polysomal RNA. Cox et al., "Plant Molecular Biology: A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The polysomal RNA was used for cDNA synthesis by the methods described below. Polysomal mRNA was then isolated as described above for corn cDNA. The quality of the RNA was assessed electrophoretically.

Following preparation of the mRNAs from various tissues as described above, selection
15 of mRNA with intact 5' ends and specific attachment of an oligonucleotide tag to the 5' end of such mRNA was performed using either a chemical or enzymatic approach. Both techniques take advantage of the presence of the "cap" structure, which characterizes the 5' end of most intact mRNAs and which comprises a guanosine generally methylated once, at the 7 position.

The chemical modification approach involves the optional elimination of the 2', 3'-cis
20 diol of the 3' terminal ribose, the oxidation of the 2', 3'-cis diol of the ribose linked to the cap of the 5' ends of the mRNAs into a dialdehyde, and the coupling of the such obtained dialdehyde to a derivatized oligonucleotide tag. Further detail regarding the chemical approaches for obtaining mRNAs having intact 5' ends are disclosed in International Application No. WO96/34981 published November 7, 1996.

25 The enzymatic approach for ligating the oligonucleotide tag to the intact 5' ends of mRNAs involves the removal of the phosphate groups present on the 5' ends of uncapped incomplete mRNAs, the subsequent decapping of mRNAs having intact 5' ends and the ligation of the phosphate present at the 5' end of the decapped mRNA to an oligonucleotide tag. Further detail regarding the enzymatic approaches for obtaining mRNAs having intact 5' ends are
30 disclosed in Dumas Milne Edwards J.B. (Doctoral Thesis of Paris VI University, Le clonage des ADNc complets: difficultés et perspectives nouvelles. Apports pour l'étude de la régulation de l'expression de la tryptophane hydroxylase de rat, 20 Dec. 1993), EP0 625572 and Kato *et al.*, *Gene* 150:243-250 (1994).

In both the chemical and the enzymatic approach, the oligonucleotide tag has a restriction enzyme site (e.g. an EcoRI site) therein to facilitate later cloning procedures. Following attachment of the oligonucleotide tag to the mRNA, the integrity of the mRNA is examined by performing a Northern blot using a probe complementary to the oligonucleotide tag.

For the mRNAs joined to oligonucleotide tags using either the chemical or the enzymatic method, first strand cDNA synthesis is performed using an oligo-dT primer with reverse transcriptase. This oligo-dT primer can contain an internal tag of at least 4 nucleotides, which can be different from one mRNA preparation to another. Methylated dCTP is used for cDNA first strand synthesis to protect the internal EcoRI sites from digestion during subsequent steps. The first strand cDNA is precipitated using isopropanol after removal of RNA by alkaline hydrolysis to eliminate residual primers.

Second strand cDNA synthesis is conducted using a DNA polymerase, such as Klenow fragment and a primer corresponding to the 5' end of the ligated oligonucleotide. The primer is typically 20-25 bases in length. Methylated dCTP is used for second strand synthesis in order to protect internal EcoRI sites in the cDNA from digestion during the cloning process.

Following second strand synthesis, the full-length cDNAs are cloned into a phagemid vector, such as pBlueScriptTM (Stratagene). The ends of the full-length cDNAs are blunted with T4 DNA polymerase (Biolabs) and the cDNA is digested with EcoRI. Since methylated dCTP is used during cDNA synthesis, the EcoRI site present in the tag is the only hemi-methylated site; hence the only site susceptible to EcoRI digestion. In some instances, to facilitate subcloning, an Hind III adapter is added to the 3' end of full-length cDNAs.

The full-length cDNAs are then size fractionated using either exclusion chromatography (AcA, Biosepra) or electrophoretic separation which yields 3 to 6 different fractions. The full-length cDNAs are then directionally cloned either into pBlueScriptTM using either the EcoRI and SmaI restriction sites or, when the Hind III adapter is present in the full-length cDNAs, the EcoRI and Hind III restriction sites. The ligation mixture is transformed, preferably by electroporation, into bacteria, which are then propagated under appropriate antibiotic selection.

Clones containing the oligonucleotide tag attached to full-length cDNAs are selected as follows.

The plasmid cDNA libraries made as described above are purified (e.g. by a column available from Qiagen). A positive selection of the tagged clones is performed as follows. Briefly, in this selection procedure, the plasmid DNA is converted to single stranded DNA using

phage F1 gene II endonuclease in combination with an exonuclease (Chang et al., *Gene* 127:95 (1993)) such as exonuclease III or T7 gene 6 exonuclease. The resulting single stranded DNA is then purified using paramagnetic beads as described by Fry et al., *Biotechniques* 13: 124 (1992). Here the single stranded DNA is hybridized with a biotinylated oligonucleotide having a
5 sequence corresponding to the 3' end of the oligonucleotide tag. Preferably, the primer has a length of 20-25 bases. Clones including a sequence complementary to the biotinylated oligonucleotide are selected by incubation with streptavidin coated magnetic beads followed by magnetic capture. After capture of the positive clones, the plasmid DNA is released from the magnetic beads and converted into double stranded DNA using a DNA polymerase such as
10 ThermoSequenase™ (obtained from Amersham Pharmacia Biotech). Alternatively, protocols such as the Gene Trapper™ kit (Gibco BRL) can be used. The double stranded DNA is then transformed, preferably by electroporation, into bacteria. The percentage of positive clones having the 5' tag oligonucleotide is typically estimated to be between 90 and 98% from dot blot analysis.

15 Following transformation, the libraries are ordered in microtiter plates and sequenced. The *Arabidopsis* library was deposited at the American Type Culture Collection on January 7, 2000 as "*E-coli* liba 010600" under the accession number PTA-1161.

EXAMPLE 2: SOUTHERN HYBRIDIZATIONS

The SDFs of the invention can be used in Southern hybridizations as described above.
20 The following describes extraction of DNA from nuclei of plant cells, digestion of the nuclear DNA and separation by length, transfer of the separated fragments to membranes, preparation of probes for hybridization, hybridization and detection of the hybridized probe.

The procedures described herein can be used to isolate related polynucleotides or for diagnostic purposes. Moderate stringency hybridization conditions, as defined above, are
25 described in the present example. These conditions result in detection of hybridization between sequences having at least 70% sequence identity. As described above, the hybridization and wash conditions can be changed to reflect the desired percentatge of sequence identity between probe and target sequences that can be detected.

In the following procedure, a probe for hybridization is produced from two PCR
30 reactions using two primers from genomic sequence of *Arabidopsis thaliana*. As described above, the particular template for generating the probe can be any desired template.

The first PCR product is assessed to validate the size of the primer to assure it is of the expected size. Then the product of the first PCR is used as a template, with the same pair

of primers used in the first PCR, in a second PCR that produces a labeled product used as the probe.

Fragments detected by hybridization, or other bands of interest, can be isolated from gels used to separate genomic DNA fragments by known methods for further purification and/or characterization.

Buffers for nuclear DNA extraction

1. 10X HB

	1000 ml	
40 mM spermidine	10.2 g	Spermine (Sigma S-2876) and spermidine (Sigma S-2501)
10 mM spermine	3.5 g	Stabilize chromatin and the nuclear membrane
0.1 M EDTA (disodium)	37.2 g	EDTA inhibits nuclease
0.1 M Tris	12.1 g	Buffer
0.8 M KCl	59.6 g	Adjusts ionic strength for stability of nuclei

Adjust pH to 9.5 with 10 N NaOH. It appears that there is a nuclease present in leaves. Use of pH 9.5 appears to inactivate this nuclease.

2. 2 M sucrose (684 g per 1000 ml)

Heat about half the final volume of water to about 50°C. Add the sucrose slowly then bring the mixture to close to final volume; stir constantly until it has dissolved. Bring the solution to volume.

3. Sarkosyl solution (lyses nuclear membranes)

	874	
N-lauroyl sarcosine (Sarkosyl)		20.0 g
0.1 M Tris		12.1 g
0.04 M EDTA (Disodium)	14.9 g	

Adjust the pH to 9.5 after all the components are dissolved and bring up to the proper volume.

4. 20% Triton X-100
80 ml Triton X-100
320 ml 1xHB (w/o β -ME and PMSF)
Prepare in advance; Triton takes some time to dissolve

A. Procedure

1. Prepare 1X "H" buffer (keep ice-cold during use)

	<u>1000 ml</u>	
10X HB		100 ml
2 M sucrose	250 ml	a non-ionic osmoticum
Water	634 ml	

Added just before use:

100 mM PMSF*	10 ml	a protease inhibitor; protects nuclear membrane proteins
β -mercaptoethanol	1 ml	inactivates nuclease by reducing disulfide bonds

*100 mM PMSF
(phenyl methyl sulfonyl fluoride, Sigma P-7626)
(add 0.0875 g to 5 ml 100% ethanol)

2. Homogenize the tissue in a blender (use 300-400 ml of 1xHB per blender). Be sure that you use 5-10 ml of HB buffer per gram of tissue. Blenders generate heat so be

sure to keep the homogenate cold. It is necessary to put the blenders in ice periodically.

3. Add the 20% Triton X-100 (25 ml per liter of homogenate) and gently stir on ice for 20 min. This lyses plastid, but not nuclear, membranes.

- 5 4. Filter the tissue suspension through several nylon filters into an ice-cold beaker. The first filtration is through a 250-micron membrane; the second is through an 85-micron membrane; the third is through a 50-micron membrane; and the fourth is through a 20-micron membrane. Use a large funnel to hold the filters. Filtration can be sped up by gently squeezing the liquid through the filters.

- 10 5. Centrifuge the filtrate at 1200 x g for 20 min. at 4°C to pellet the nuclei.

6. Discard the dark green supernatant. The pellet will have several layers to it. One is starch; it is white and gritty. The nuclei are gray and soft. In the early steps, there may be a dark green and somewhat viscous layer of chloroplasts.

15 Wash the pellets in about 25 ml cold H buffer (with Triton X-100) and resuspend by swirling gently and pipetting. After the pellets are resuspended.

Pellet the nuclei again at 1200 - 1300 x g. Discard the supernatant.

20 Repeat the wash 3-4 times until the supernatant has changed from a dark green to a pale green. This usually happens after 3 or 4 resuspensions. At this point, the pellet is typically grayish white and very slippery. The Triton X-100 in these repeated steps helps to destroy the chloroplasts and mitochondria that contaminate the prep.

Resuspend the nuclei for a final time in a total of 15 ml of H buffer and transfer the suspension to a sterile 125 ml Erlenmeyer flask.

7. Add 15 ml, dropwise, cold 2% Sarkosyl, 0.1 M Tris, 0.04 M EDTA solution (pH 9.5) while swirling gently. This lyses the nuclei. The solution will become very viscous.

8. Add 30 grams of CsCl and gently swirl at room temperature until the CsCl is in solution. The mixture will be gray, white and viscous.
9. Centrifuge the solution at 11,400 x g at 4°C for at least 30 min. The longer this spin is, the firmer the protein pellicle.
- 5 10. The result is typically a clear green supernatant over a white pellet, and (perhaps) under a protein pellicle. Carefully remove the solution under the protein pellicle and above the pellet. Determine the density of the solution by weighing 1 ml of solution and add CsCl if necessary to bring to 1.57 g/ml. The solution contains dissolved
10 solids (sucrose etc) and the refractive index alone will not be an accurate guide to CsCl concentration.
11. Add 20 µl of 10 mg/ml EtBr per ml of solution.
12. Centrifuge at 184,000 x g for 16 to 20 hours in a fixed-angle rotor.
13. Remove the dark red supernatant that is at the top of the tube with a plastic transfer
15 pipette and discard. Carefully remove the DNA band with another transfer pipette. The DNA band is usually visible in room light; otherwise, use a long wave UV light to locate the band.
14. Extract the ethidium bromide with isopropanol saturated with water and salt. Once
20 the solution is clear, extract at least two more times to ensure that all of the EtBr is gone. Be very gentle, as it is very easy to shear the DNA at this step. This extraction may take a while because the DNA solution tends to be very viscous. If the solution is too viscous, dilute it with TE.
15. Dialyze the DNA for at least two days against several changes (at least three times) of TE (10 mM Tris, 1mM EDTA, pH 8) to remove the cesium chloride.

16. Remove the dialyzed DNA from the tubing. If the dialyzed DNA solution contains a lot of debris, centrifuge the DNA solution at least at 2500 x g for 10 min. and carefully transfer the clear supernatant to a new tube. Read the A260 concentration of the DNA.

5 17. Assess the quality of the DNA by agarose gel electrophoresis (1% agarose gel) of the DNA. Load 50 ng and 100 ng (based on the OD reading) and compare it with known and good quality DNA. Undigested lambda DNA and a lambda-HindIII-digested DNA are good molecular weight makers.

Protocol for Digestion of Genomic DNA

Protocol:

10 1. The relative amounts of DNA for different crop plants that provide approximately a balanced number of genome equivalent is given in Table 3. Note that due to the size of the wheat genome, wheat DNA will be underrepresented. Lambda DNA provides a useful control for complete digestion.

15 2. Precipitate the DNA by adding 3 volumes of 100% ethanol. Incubate at -20°C for at least two hours. Yeast DNA can be purchased and made up at the necessary concentration, therefore no precipitation is necessary for yeast DNA.

20 3. Centrifuge the solution at 11,400 x g for 20 min. Decant the ethanol carefully (be careful not to disturb the pellet). Be sure that the residual ethanol is completely removed either by vacuum desiccation or by carefully wiping the sides of the tubes with a clean tissue.

4. Resuspend the pellet in an appropriate volume of water. Be sure the pellet is fully resuspended before proceeding to the next step. This may take about 30 min.

25 5. Add the appropriate volume of 10X reaction buffer provided by the manufacturer of the restriction enzyme to the resuspended DNA followed by the appropriate volume of enzymes. Be sure to mix it properly by slowly swirling the tubes.

6. Set-up the lambda digestion-control for each DNA that you are digesting.
7. Incubate both the experimental and lambda digests overnight at 37°C. Spin down condensation in a microfuge before proceeding.
8. After digestion, add 2 µl of loading dye (typically 0.25% bromophenol blue, 0.25% xylene cyanol in 15% Ficoll or 30% glycerol) to the lambda-control digests and load in 1% TPE-agarose gel (TPE is 90 mM Tris-phosphate, 2 mM EDTA, pH 8). If the lambda DNA in the lambda control digests are completely digested, proceed with the precipitation of the genomic DNA in the digests.
9. Precipitate the digested DNA by adding 3 volumes of 100% ethanol and incubating in -20°C for at least 2 hours (preferably overnight).

EXCEPTION: *Arabidopsis* and yeast DNA are digested in an appropriate volume; they don't have to be precipitated.

10. Resuspend the DNA in an appropriate volume of TE (e.g., 22 µl x 50 blots = 1100 µl) and an appropriate volume of 10X loading dye (e.g., 2.4 µl x 50 blots = 120 µl). Be careful in pipetting the loading dye - it is viscous. Be sure you are pipetting the correct volume.

Table 3

Some guide points in digesting genomic DNA.

Species	Genome Size	Size Relative to Arabidopsis	Genome Equivalent to 2 µg Arabidopsis DNA	Amount of DNA per blot
Arabidopsis	120 Mb	1X	1X	2 µg
Brassica	1,100 Mb	9.2X	0.54X	10 µg
Corn	2,800 Mb	23.3X	0.43X	20 µg

879

Cotton	2,300 Mb	19.2X	0.52X	20 µg
Oat	11,300 Mb	94X	0.11X	20 µg
Rice	400 Mb	3.3X	0.75X	5 µg
Soybean	1,100 Mb	9.2X	0.54X	10 µg
Sugarbeet	758 Mb	6.3X	0.8X	10 µg
Sweetclover	1,100 Mb	9.2X	0.54X	10 µg
Wheat	16,000 Mb	133X	0.08X	20 µg
Yeast	15 Mb	0.12X	1X	0.25 µg

Protocol for Southern Blot Analysis

The digested DNA samples are electrophoresed in 1% agarose gels in 1x TPE buffer. Low voltage; overnight separations are preferred. The gels are stained with EtBr and photographed.

1. For blotting the gels, first incubate the gel in 0.25 N HCl (with gentle shaking) for about 15 min.
2. Then briefly rinse with water. The DNA is denatured by 2 incubations. Incubate (with shaking) in 0.5 M NaOH in 1.5 M NaCl for 15 min.
3. The gel is then briefly rinsed in water and neutralized by incubating twice (with shaking) in 1.5 M Tris pH 7.5 in 1.5 M NaCl for 15 min.
4. A nylon membrane is prepared by soaking it in water for at least 5 min, then in 6X SSC for at least 15 min. before use. (20x SSC is 175.3 g NaCl, 88.2 g sodium citrate per liter, adjusted to pH 7.0.)
5. The nylon membrane is placed on top of the gel and all bubbles in between are removed. The DNA is blotted from the gel to the membrane using an absorbent medium, such as paper toweling and 6x SCC buffer. After the transfer, the membrane may be lightly brushed with a gloved hand to remove any agarose sticking to the surface.

6. The DNA is then fixed to the membrane by UV crosslinking and baking at 80°C. The membrane is stored at 4°C until use.

B. Protocol for PCR Amplification of Genomic Fragments in Arabidopsis

Amplification procedures:

- 5 1. Mix the following in a 0.20 ml PCR tube or 96-well PCR plate:

Volume	Stock	Final Amount or Conc.
0.5 µl	~ 10 ng/µl genomic DNA ¹	5 ng
2.5 µl	10X PCR buffer	20 mM Tris, 50 mM KCl
0.75 µl	50 mM MgCl ₂	1.5 mM
1 µl	10 pmol/µl Primer 1 (Forward)	10 pmol
1 µl	10 pmol/µl Primer 2 (Reverse)	10 pmol
0.5 µl	5 mM dNTPs	0.1 mM
0.1 µl	5 units/µl Platinum Taq™ (Life Technologies, Gaithersburg, MD) DNA Polymerase	1 units
(to 25 µl)	Water	

2. The template DNA is amplified using a Perkin Elmer 9700 PCR machine:

¹ Arabidopsis DNA is used in the present experiment, but the procedure is a general one.

- 1) 94°C for 10 min. followed by

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
94 °C - 30 sec 62 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 58 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 53 °C - 30 sec 72 °C - 3 min

- 5) 72°C for 7 min. Then the reactions are stopped by chilling to 4°C.

The procedure can be adapted to a multi-well format if necessary.

Quantification and Dilution of PCR Products:

- 5 1. The product of the PCR is analyzed by electrophoresis in a 1% agarose gel. A linearized plasmid DNA can be used as a quantification standard (usually at 50, 100, 200, and 400 ng). These will be used as references to approximate the amount of PCR products. HindIII-digested Lambda DNA is useful as a molecular weight marker. The gel can be run fairly quickly; e.g., at 100 volts. The standard gel is examined to determine that the size of the PCR products is consistent with the expected size and if there are significant extra bands or smeary products in the PCR reactions.
2. The amounts of PCR products can be estimated on the basis of the plasmid standard.
3. For the small number of reactions that produce extraneous bands, a small amount of DNA from bands with the correct size can be isolated by dipping a sterile 10-μl tip into the band while viewing through a UV Transilluminator. The small amount of agarose gel (with the DNA fragment) is used in the labeling reaction.

C. Protocol for PCR-DIG-Labeling of DNA

Solutions:

Reagents in PCR reactions (diluted PCR products, 10X PCR Buffer, 50 mM MgCl₂, 5 U/μl Platinum Taq Polymerase, and the primers)

10X dNTP + DIG-11-dUTP [1:5]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.65 mM dTTP, 0.35 mM DIG-11-dUTP)

5 10X dNTP + DIG-11-dUTP [1:10]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.81 mM dTTP, 0.19 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:15]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.875 mM dTTP, 0.125 mM DIG-11-dUTP)

TE buffer (10 mM Tris, 1 mM EDTA, pH 8)

10 Maleate buffer: In 700 ml of deionized distilled water, dissolve 11.61 g maleic acid and 8.77 g NaCl. Add NaOH to adjust the pH to 7.5. Bring the volume to 1 L. Stir for 15 min. and sterilize.

15 10% blocking solution: In 80 ml deionized distilled water, dissolve 1.16g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, Cat. no. 1096176). Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

1% blocking solution: Dilute the 10% stock to 1% using the maleate buffer.

20 Buffer 3 (100 mM Tris, 100 mM NaCl, 50 mM MgCl₂, pH9.5). Prepared from autoclaved solutions of 1M Tris pH 9.5, 5 M NaCl, and 1 M MgCl₂ in autoclaved distilled water.

Procedure:

1. PCR reactions are performed in 25 µl volumes containing:

PCR buffer	1X
MgCl ₂	1.5 mM
10X dNTP + DIG-11-dUTP	1X (please see the note below)
Platinum Taq™ Polymerase	1 unit
10 pg probe DNA	
10 pmol primer 1	

Note:**Use for:**

10X dNTP + DIG-11-dUTP (1:5)	< 1 kb
10X dNTP + DIG-11-dUTP (1:10)	1 kb to 1.8 kb
10X dNTP + DIG-11-dUTP (1:15)	> 1.8 kb

2. The PCR reaction uses the following amplification cycles:

- 1) 94°C for 10 min.

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
95°C - 30 sec 61°C - 1 min 73°C - 5 min	95°C - 30 sec 59°C - 1 min 75°C - 5 min	95°C - 30 sec 51°C - 1 min 73°C - 5 min

- 5) 72°C for 8 min. The reactions are terminated by chilling to 4°C (hold).

3. The products are analyzed by electrophoresis- in a 1% agarose gel, comparing to an aliquot of the unlabelled probe starting material.
4. The amount of DIG-labeled probe is determined as follows:

Make serial dilutions of the diluted control DNA in dilution buffer (TE: 10 mM Tris and 1 mM EDTA, pH 8) as shown in the following table:

DIG-labeled control DNA starting conc.	Stepwise Dilution	Final Conc. (Dilution Name)
5 ng/ μ l	1 μ l in 49 μ l TE	100 pg/ μ l (A)
100 pg/ μ l (A)	25 μ l in 25 μ l TE	50 pg/ μ l (B)
50 pg/ μ l (B)	25 μ l in 25 μ l TE	25 pg/ μ l (C)
25 pg/ μ l (C)	20 μ l in 30 μ l TE	10 pg/ μ l (D)

- a. Serial dilutions of a DIG-labeled standard DNA ranging from 100 pg to 10 pg are spotted onto a positively charged nylon membrane, marking the membrane lightly with a pencil to identify each dilution.
- b. Serial dilutions (e.g., 1:50, 1:2500, 1:10,000) of the newly labeled DNA probe are spotted.
- c. The membrane is fixed by UV crosslinking.
- d. The membrane is wetted with a small amount of maleate buffer and then incubated in 1% blocking solution for 15 min at room temp.
- e. The labeled DNA is then detected using alkaline phosphatase conjugated anti-DIG antibody (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) and an NBT substrate according to the manufacture's instruction.
- f. Spot intensities of the control and experimental dilutions are then compared to estimate the concentration of the PCR-DIG-labeled probe.

D. Prehybridization and Hybridization of Southern BlotsSolutions:

100% Formamide purchased from Gibco

20X SSC (1X = 0.15 M NaCl, 0.015 M Na₃citrate)

5 per L: 175 g NaCl
 87.5 g Na₃citrate·2H₂O

20% Sarkosyl (N-lauroyl-sarcosine)

20% SDS (sodium dodecyl sulphate)

10 10% Blocking Reagent: In 80 ml deionized distilled water, dissolve 1.16 g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder. Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

Prehybridization Mix:

Final Concentration	Components	Volume (per 100 ml)	Stock
50%	Formamide	50 ml	100%
5X	SSC	25 ml	20X
0.1%	Sarkosyl	0.5 ml	20%
0.02%	SDS	0.1 ml	20%
2%	Blocking Reagent	20 ml	10%
	Water	4.4 ml	

General Procedures:

- 15 1. Place the blot in a heat-sealable plastic bag and add an appropriate volume of prehybridization solution (30 ml/100cm²) at room temperature. Seal the bag with a heat sealer, avoiding bubbles as much as possible. Lay down the bags in a large plastic tray (one tray can accommodate at least 4–5 bags). Ensure that the bags are

lying flat in the tray so that the prehybridization solution is evenly distributed throughout the bag. Incubate the blot for at least 2 hours with gentle agitation using a waver shaker.

2. Denature DIG-labeled DNA probe by incubating for 10 min. at 98°C using the PCR machine and immediately cool it to 4°C.

3. Add probe to prehybridization solution (25 ng/ml; 30 ml = 750 ng total probe) and mix well but avoid foaming. Bubbles may lead to background.

4. Pour off the prehybridization solution from the hybridization bags and add new prehybridization and probe solution mixture to the bags containing the membrane.

5. Incubate with gentle agitation for at least 16 hours.

6. Proceed to medium stringency post-hybridization wash:

Three times for 20 min. each with gentle agitation using 1X SSC, 1% SDS at 60°C.

All wash solutions must be prewarmed to 60°C. Use about 100 ml of wash solution per membrane.

To avoid background keep the membranes fully submerged to avoid drying in spots; agitate sufficiently to avoid having membranes stick to one another.

7. After the wash, proceed to immunological detection and CSPD development.

E. Procedure for Immunological Detection with CSPD

Solutions:

Buffer 1: Maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl; adjusted to pH 7.5 with NaOH)

Washing buffer: Maleic acid buffer with 0.3% (v/v) Tween 20.

Blocking stock solution 10% blocking reagent in buffer 1. Dissolve (10X concentration): blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, cat. no. 1096176) by constantly stirring on a 65°C heating block or heat in a microwave, autoclave and store at 4°C.

Buffer 2

(1X blocking solution): Dilute the stock solution 1:10 in Buffer 1.

Detection buffer: 0.1 M Tris, 0.1 M NaCl, pH 9.5

10 Procedure:

1. After the post-hybridization wash the blots are briefly rinsed (1-5 min.) in the maleate washing buffer with gentle shaking.
2. Then the membranes are incubated for 30 min. in Buffer 2 with gentle shaking.
3. Anti-DIG-AP conjugate (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) at 75 mU/ml (1:10,000) in Buffer 2 is used for detection. 75 ml of solution can be used for 3 blots.
4. The membrane is incubated for 30 min. in the antibody solution with gentle shaking.
5. The membrane are washed twice in washing buffer with gentle shaking. About 250 mls is used per wash for 3 blots.
6. The blots are equilibrated for 2-5 min in 60 ml detection buffer.
7. Dilute CSPD (1:200) in detection buffer. (This can be prepared ahead of time and stored in the dark at 4°C).

The following steps must be done individually. Bags (one for detection and one for exposure) are generally cut and ready before doing the following steps.

8. The blot is carefully removed from the detection buffer and excess liquid removed without drying the membrane. The blot is immediately placed in a bag and 1.5 ml of CSPD solution is added. The CSPD solution can be spread over the membrane. Bubbles present at the edge and on the surface of the blot are typically removed by gentle rubbing. The membrane is incubated for 5 min. in CSPD solution.
9. Excess liquid is removed and the membrane is blotted briefly (DNA side up) on Whatman 3MM paper. Do not let the membrane dry completely.
10. Seal the damp membrane in a hybridization bag and incubate for 10 min at 37°C to enhance the luminescent reaction.
11. Expose for 2 hours at room temperature to X-ray film. Multiple exposures can be taken. Luminescence continues for at least 24 hours and signal intensity increases during the first hours.

Example 3: Transformation of Carrot Cells

Transformation of plant cells can be accomplished by a number of methods, as described above. Similarly, a number of plant genera can be regenerated from tissue culture following transformation. Transformation and regeneration of carrot cells as described herein is illustrative.

Single cell suspension cultures of carrot (*Daucus carota*) cells are established from hypocotyls of cultivar Early Nantes in B₅ growth medium (O.L. Gamborg et al., *Plant Physiol.* 45:372 (1970)) plus 2,4-D and 15 mM CaCl₂ (B₅-44 medium) by methods known in the art. The suspension cultures are subcultured by adding 10 ml of the suspension culture to 40 ml of B₅-44 medium in 250 ml flasks every 7 days and are maintained in a shaker at 150 rpm at 27 °C in the dark.

The suspension culture cells are transformed with exogenous DNA as described by Z. Chen et al. *Plant Mol. Bio.* 36:163 (1998). Briefly, 4-days post-subculture cells are incubated with cell wall digestion solution containing 0.4 M sorbitol, 2% driselase, 5mM MES (2-[N-

Morpholino] ethanesulfonic acid) pH 5.0 for 5 hours. The digested cells are pelleted gently at 60 xg for 5 min. and washed twice in W5 solution containing 154 mM NaCl, 5 mM KCl, 125 mM CaCl₂ and 5mM glucose, pH 6.0. The protoplasts are suspended in MC solution containing 5 mM MES, 20 mM CaCl₂, 0.5 M mannitol, pH 5.7 and the protoplast density is adjusted to about 4 x 10⁶ protoplasts per ml.

15-60 µg of plasmid DNA is mixed with 0.9 ml of protoplasts. The resulting suspension is mixed with 40% polyethylene glycol (MW 8000, PEG 8000), by gentle inversion a few times at room temperature for 5 to 25 min. Protoplast culture medium known in the art is added into the PEG-DNA-protoplast mixture. Protoplasts are incubated in the culture medium for 24 hour to 5 days and cell extracts can be used for assay of transient expression of the introduced gene. Alternatively, transformed cells can be used to produce transgenic callus, which in turn can be used to produce transgenic plants, by methods known in the art. See, for example, Nomura and Komamine, *Plt. Phys.* 79:988-991 (1985), *Identification and Isolation of Single Cells that Produce Somatic Embryos in Carrot Suspension Cultures*.

An additional deposit, PTA-1411, of an *E. coli* Library, *E. coli*LibA021800, was made at the American Type Culture Collection in Manassas, Virginia, USA on February 22, 2000 to meet the requirements of Budapest Treaty for the international recognition of the deposit of microorganisms. This deposit was assigned ATCC accession no. PTA-1411.

The invention being thus described, it will be apparent to one of ordinary skill in the art that various modifications of the materials and methods for practicing the invention can be made. Such modifications are to be considered within the scope of the invention as defined by the following claims.

Each of the references from the patent and periodical literature cited herein is hereby expressly incorporated in its entirety by such citation.

TABLE 1

	>1297184	/22656			
5	len =	1421	nex =	3	
	Term	10090	9717	-	0
	Intr	10506	10184	-	0
	Init	11137	10900	-	0
10	>1297184	/38841			
	len =	1470	nex =	4	
15	Term	14341	13880	-	0
	Intr	14529	14477	-	0
	Intr	14673	14624	-	0
	Init	15349	15056	-	0
20	>1297184	/40037			
	len =	1735	nex =	3	
	Init	16472	16883	+	0
25	Intr	17095	17382	+	0
	Term	17730	18206	+	0
	>1297184	/37635			
30	len =	3070	nex =	11	
	Init	23715	23788	+	0
	Intr	24275	24361	+	0
	Intr	24477	24554	+	0
35	Intr	24641	24834	+	0
	Intr	24949	25090	+	0
	Intr	25275	25355	+	0
	Intr	25618	25746	+	0
	Intr	25852	25929	+	0
40	Intr	26008	26079	+	0
	Intr	26239	26319	+	0
	Term	26416	26618	+	0
	>1402874	/32010			
45	len =	1171	nex =	4	
	Init	65717	66071	+	0
	Intr	66169	66290	+	0
50	Intr	66363	66515	+	0
	Term	66600	66887	+	0
	>1402874	/16813			
55	len =	753	nex =	3	
	Init	78870	78982	+	0
	Intr	79066	79242	+	0
	Term	79311	79622	+	0
60					

	>1402874	/41074		
	len =	850	nex =	3
5	Init	80005	80250	+
	Intr	80334	80486	+
	Term	80573	80854	+
				0
				0
10	>1532162	/42644		
	len =	1353	nex =	4
	Init	10117	10395	+
	Intr	10519	10718	+
15	Intr	10809	11038	+
	Term	11112	11469	+
				0
				0
	>1532162	/156172		
20	len =	1286	nex =	4
	Init	10232	10395	+
	Intr	10519	10718	+
	Intr	10809	11038	+
25	Term	11112	11517	+
				0
				0
	>1532162	/1415		
30	len =	649	nex =	2
	Init	11955	12231	+
	Term	12334	12603	+
				0
				0
	>1532162	/32937		
35	len =	738	nex =	2
	Init	11982	12231	+
	Term	12334	12719	+
				0
				0
40	>1532162	/25057		
	len =	2230	nex =	9
45	Init	21905	22035	+
	Intr	22473	22603	+
	Intr	22691	22865	+
	Intr	22971	23036	+
	Intr	23189	23323	+
50	Intr	23415	23494	+
	Intr	23568	23670	+
	Intr	23758	23816	+
	Term	23900	24131	+
				0
				0
55	>1532162	/20800		
	len =	2174	nex =	9
	Init	22161	22221	+
60	Intr	22473	22603	+
				0
				0

				892	
	Intr	22691	22865	+	0
	Intr	22971	23036	+	0
	Intr	23189	23323	+	0
	Intr	23415	23494	+	0
5	Intr	23568	23670	+	0
	Intr	23758	23816	+	0
	Term	23900	24131	+	0
	>1532162	/33957			
10	len =	2175	nex =	9	
	Init	22161	22221	+	0
	Intr	22473	22603	+	0
15	Intr	22691	22865	+	0
	Intr	22971	23036	+	0
	Intr	23189	23323	+	0
	Intr	23415	23494	+	0
	Intr	23568	23670	+	0
20	Intr	23758	23816	+	0
	Term	23900	24132	+	0
	>1532162	/154048			
25	len =	638	nex =	1	
	Sngl	28309	28946	+	0
	>1532162	/15529			
30	len =	2414	nex =	1	
	Sngl	40592	41742	+	0
	>1532162	/39051			
35	len =	2127	nex =	1	
	Sngl	45263	44433	-	0
40	>1532162	/15968			
	len =	191	nex =	1	
45	Sngl	47765	47575	-	0
	>1532162	/29991			
	len =	2374	nex =	1	
50	Sngl	48703	47562	-	0
	>1532162	/6135			
55	len =	2370	nex =	1	
	Sngl	48703	47566	-	0
	>1532162	/26043			
60					

893

	len =	1243	nex =	3	
	Term	50762	50243	-	0
	Intr	51173	51108	-	0
5	Init	51485	51286	-	0
	>1532162	/14942			
10	len =	1966	nex =	3	
	Init	52773	52831	+	0
	Intr	52920	53095	+	0
	Term	53208	53601	+	0
15	>1532162	/98909			
	len =	1886	nex =	3	
	Init	52773	52831	+	0
20	Intr	52920	53095	+	0
	Term	53208	53533	+	0
	>1532162	/28026			
25	len =	1675	nex =	4	
	Term	58063	57720	-	0
	Intr	58192	58151	-	0
	Intr	58514	58430	-	0
30	Init	59394	58958	-	0
	>1707006	/26007			
35	len =	359	nex =	2	
	Init	22636	22811	+	0
	Term	22887	22994	+	0
40	>1707006	/26506			
	len =	2124	nex =	6	
	Init	22665	22811	+	0
	Intr	22887	22993	+	0
45	Intr	23509	23571	+	0
	Intr	23932	23963	+	0
	Intr	24280	24374	+	0
	Term	24494	24773	+	0
50	>1707006	/40748			
	len =	3502	nex =	7	
	Init	50053	50271	+	0
55	Intr	50378	50467	+	0
	Intr	51366	51425	+	0
	Intr	51531	51570	+	0
	Intr	51655	51748	+	0
	Intr	51858	51917	+	0
60	Term	52003	52094	+	0

	>1707006	/125567			
5	len =	1097	nex =	2	
	Term	54221	53867	-	0
	Init	54963	54743	-	0
10	>1707006	/152227			
	len =	430	nex =	1	
	Sngl	55477	55054	-	0
15	>1707006	/38063			
	len =	1598	nex =	2	
20	Term	54226	53885	-	0
	Init	55482	54743	-	0
	>1707006	/10375			
25	len =	744	nex =	2	
	Init	58320	58692	+	0
	Term	58784	59063	+	0
30	>1707006	/10617			
	len =	815	nex =	3	
	Term	62856	62577	-	0
	Intr	63141	62943	-	0
35	Init	63391	63248	-	0
	>1707006	/1711			
40	len =	801	nex =	3	
	Term	62856	62591	-	0
	Intr	63141	62943	-	0
	Init	63391	63248	-	0
45	>1707006	/29818			
	len =	760	nex =	3	
50	Term	62856	62632	-	0
	Intr	63141	62943	-	0
	Init	63391	63248	-	0
	>1707006	/40627			
55	len =	2499	nex =	10	
	Init	67998	68163	+	0
	Intr	68353	68506	+	0
	Intr	68592	68875	+	0
60	Intr	68968	69064	+	0

				895	
	Intr	69314	69419	+	0
	Intr	69514	69596	+	0
	Intr	69689	69834	+	0
	Intr	69966	70071	+	0
5	Intr	70216	70275	+	0
	Term	70361	70496	+	0
	>1707006	/101081			
10	len =	976	nex =	1	
	Sngl	75101	74126	-	0
	>1785673	/23693			
15	len =	622	nex =	1	
	Sngl	31275	31896	+	0
20	>1871173	/38610			
	len =	2054	nex =	8	
	Init	12146	12237	+	0
25	Intr	12284	12483	+	0
	Intr	12573	12704	+	0
	Intr	12790	12866	+	0
	Intr	12970	13105	+	0
	Intr	13186	13326	+	0
30	Intr	13424	13482	+	0
	Term	13584	13650	+	0
	>1871173	/10969			
35	len =	1038	nex =	2	
	Term	49483	49183	-	0
	Init	50220	49576	-	0
40	>1871173	/16493			
	len =	3870	nex =	7	
	Init	50368	50540	+	0
45	Intr	50630	50680	+	0
	Intr	51247	51298	+	0
	Intr	51370	51428	+	0
	Intr	51546	51591	+	0
	Intr	52100	52233	+	0
50	Term	54023	54234	+	0
	>1877523	/96448			
	len =	822	nex =	4	
55	Init	105150	105228	+	0
	Intr	105315	105428	+	0
	Intr	105500	105535	+	0
	Term	105630	105953	+	0
60					

	>1877523	/2677		
	len =	670	nex =	1
5	Sngl	21255	20592	- 0
	>1877523	/1693		
	len =	710	nex =	2
10	Term	21139	20646	- 0
	Init	21355	21247	- 0
	>1877523	/40042		
15	len =	2369	nex =	8
	Term	38060	37972	- 0
	Intr	38794	38656	- 0
20	Intr	38997	38927	- 0
	Intr	39148	39104	- 0
	Intr	39328	39239	- 0
	Intr	39526	39438	- 0
	Intr	39689	39614	- 0
25	Init	40034	39798	- 0
	>1877523	/35733		
	len =	1245	nex =	3
30	Term	53914	53583	- 0
	Intr	54335	54159	- 0
	Init	54827	54574	- 0
	>1877523	/34291		
	len =	809	nex =	1
	Sngl	61137	61281	+ 0
40	>1877523	/2979		
	len =	766	nex =	2
	Init	61146	61281	+ 0
45	Term	61565	61911	+ 0
	>1931636	/93598		
	len =	1589	nex =	3
	Init	111855	112746	+ 0
	Intr	112845	112949	+ 0
	Term	113156	113443	+ 0
55	>1931636	/40765		
	len =	1821	nex =	3
60	Term	50015	49475	- 0

				897	
	Intr	50253	50130	-	0
	Init	51295	50557	-	0
5	>1931636	/20637			
	len =	644	nex =	1	
	Sngl	63596	62953	-	0
10	>1931636	/14648			
	len =	503	nex =	1	
	Sngl	97733	97231	-	0
15	>1946354	/1391			
	len =	4584	nex =	11	
20	Term	12119	11739	-	0
	Intr	12281	12213	-	0
	Intr	12535	12455	-	0
	Intr	12756	12682	-	0
	Intr	13005	12873	-	0
25	Intr	13304	13257	-	0
	Intr	13613	13401	-	0
	Intr	13994	13833	-	0
	Intr	14593	14363	-	0
	Intr	15009	14680	-	0
30	Init	15456	15157	-	0
	>1946354	/7619			
	len =	939	nex =	2	
35	Init	31875	32384	+	0
	Term	32537	32813	+	0
40	>1946354	/34999			
	len =	1078	nex =	2	
	Init	33182	33416	+	0
	Term	33743	34259	+	0
45	>1946354	/39560			
	len =	674	nex =	1	
50	Sngl	41592	42265	+	0
	>1946354	/41046			
	len =	730	nex =	1	
55	Sngl	57609	58323	+	0
	>1946354	/1820			
60	len =	1190	nex =	2	

898

	Term	7729	6909	-	0
	Init	8098	7816	-	0
5	>1946354	/22671			
	len =	1583	nex =	2	
	Init	83167	83385	+	0
10	Term	83523	83614	+	0
	>2062153	/38051			
	len =	1491	nex =	3	
15	Term	15272	14834	-	0
	Intr	15841	15386	-	0
	Init	16324	16026	-	0
20	>2062153	/119458			
	len =	1513	nex =	1	
	Sngl	16220	16026	-	0
25	>2062153	/157474			
	len =	1497	nex =	3	
30	Term	15272	14858	-	0
	Intr	15841	15386	-	0
	Init	16220	16026	-	0
	>2062153	/30056			
35	len =	1520	nex =	3	
	Term	15272	14836	-	0
	Intr	15841	15386	-	0
40	Init	16220	16026	-	0
	>2062153	/42777			
	len =	1450	nex =	2	
45	Term	24390	23947	-	0
	Init	25283	24512	-	0
	>2062153	/6448			
50	len =	1481	nex =	3	
	Term	24390	23947	-	0
	Intr	24955	24512	-	0
55	Init	25427	25053	-	0
	>2062153	/12715			
	len =	1976	nex =	3	
60					

				899	
	Term	55961	55118	-	0
	Intr	56262	56051	-	0
	Init	57093	56580	-	0
5	>2062153	/30003			
	len =	2057	nex =	3	
	Init	7382	7843	+	0
10	Intr	7929	8378	+	0
	Term	8469	8866	+	0
	>2062153	/32293			
15	len =	790	nex =	1	
	Sngl	69530	68750	-	0
	>2062153	/29750			
20	len =	2112	nex =	3	
	Init	76786	77284	+	0
	Intr	77663	77774	+	0
25	Term	77921	78394	+	0
	>2088638	/9398			
	len =	616	nex =	1	
30	Sngl	103573	102958	-	0
	>2088638	/6732			
35	len =	1632	nex =	2	
	Term	17390	16530	-	0
	Init	18161	17822	-	0
40	>2088638	/39048			
	len =	2533	nex =	7	
	Init	24452	24782	+	0
45	Intr	25154	25378	+	0
	Intr	25457	25551	+	0
	Intr	25633	25822	+	0
	Intr	25917	26041	+	0
	Intr	26186	26401	+	0
50	Term	26486	26984	+	0
	>2088638	/33701			
	len =	1515	nex =	3	
55	Term	32027	31685	-	0
	Intr	32312	32109	-	0
	Init	32802	32388	-	0
60	>2088638	/15207			

900

	len =	2110	nex =	4	
	Term	50426	50181	-	0
5	Intr	50656	50514	-	0
	Intr	51540	51487	-	0
	Init	52290	52051	-	0
	>2088638 /5504				
10	len =	2820	nex =	10	
	Term	52859	52504	-	0
	Intr	53066	52943	-	0
15	Intr	53260	53159	-	0
	Intr	53424	53356	-	0
	Intr	53674	53567	-	0
	Intr	53905	53851	-	0
	Intr	54431	54301	-	0
20	Intr	54618	54544	-	0
	Intr	54880	54803	-	0
	Init	55058	54973	-	0
	>2088638 /35056				
25	len =	1510	nex =	1	
	Sngl	70686	69178	-	0
30	>2088638 /32440				
	len =	919	nex =	3	
	Init	80756	80853	+	0
35	Intr	81026	81170	+	0
	Term	81258	81674	+	0
	>2088638 /5046				
40	len =	1647	nex =	3	
	Init	95145	95511	+	0
	Intr	95860	96013	+	0
	Term	96327	96791	+	0
45	>2098816 /31252				
	len =	704	nex =	1	
50	Sngl	35121	35824	+	0
	>2098816 /15292				
	len =	1279	nex =	1	
55	Sngl	39507	40333	+	0
	>2098816 /36730				
60	len =	2135	nex =	9	

901

	Init	43827	44181	+	0
	Intr	44267	44314	+	0
	Intr	44406	44582	+	0
5	Intr	44668	44818	+	0
	Intr	44908	44994	+	0
	Intr	45079	45203	+	0
	Intr	45282	45400	+	0
	Intr	45483	45594	+	0
10	Term	45685	45961	+	0
>2098816		/8716			
15	len =	1090	nex =	5	
	Init	44941	44994	+	0
	Intr	45079	45203	+	0
	Intr	45282	45400	+	0
	Intr	45483	45594	+	0
20	Term	45685	46007	+	0
>2098816		/36216			
25	len =	2338	nex =	6	
	Init	58990	59535	+	0
	Intr	59663	59944	+	0
	Intr	60031	60178	+	0
	Intr	60282	60367	+	0
30	Intr	60894	60971	+	0
	Term	61070	61327	+	0
>2098816		/42713			
35	len =	2280	nex =	6	
	Init	59046	59535	+	0
	Intr	59663	59944	+	0
	Intr	60031	60178	+	0
40	Intr	60282	60367	+	0
	Intr	60894	60971	+	0
	Term	61070	61325	+	0
>2098816		/36286			
45	len =	643	nex =	1	
	Sngl	6052	5410	-	0
50	>2098816	/38820			
	len =	756	nex =	1	
	Sngl	6188	5433	-	0
55	>2098816	/38170			
	len =	2445	nex =	6	
60	Term	63428	62916	-	0

				902
	Intr	63750	63522	- 0
	Intr	63933	63894	- 0
	Intr	64507	64381	- 0
	Intr	64935	64803	- 0
5	Init	65360	65060	- 0
	>2098816 /40254			
10	len =	628	nex =	1
	Sngl	69008	68744	- 0
	>2098816 /17126			
15	len =	811	nex =	2
	Term	69008	68666	- 0
	Init	69476	69328	- 0
20	>2098816 /122497			
	len =	359	nex =	1
25	Sngl	70110	69752	- 0
	>2098816 /36543			
	len =	1173	nex =	1
30	Sngl	77771	76602	- 0
	>2098816 /17357			
35	len =	2350	nex =	6
	Init	88159	88663	+ 0
	Intr	88942	89027	+ 0
	Intr	89118	89341	+ 0
	Intr	89580	89646	+ 0
40	Intr	90016	90126	+ 0
	Term	90323	90506	+ 0
	>2098816 /31770			
45	len =	1183	nex =	3
	Term	90853	90569	- 0
	Intr	91066	90933	- 0
	Init	91751	91546	- 0
50	>2104523 /21952			
	len =	2710	nex =	2
55	Term	71964	70632	- 0
	Init	73339	72021	- 0
	>2104523 /34676			
60	len =	3970	nex =	12

903

	Term	76257	75990	-	0
	Intr	76489	76350	-	0
	Intr	77518	77169	-	0
5	Intr	77737	77610	-	0
	Intr	77956	77828	-	0
	Intr	78109	78031	-	0
	Intr	78360	78197	-	0
	Intr	78636	78451	-	0
10	Intr	78894	78763	-	0
	Intr	79089	78998	-	0
	Intr	79279	79180	-	0
	Init	79954	79652	-	0
15	>2160132	/9002			
	len =	1994	nex =	4	
	Term	38308	37073	-	0
20	Intr	38529	38400	-	0
	Intr	38751	38614	-	0
	Init	39066	38884	-	0
	>2160132	/18804			
25	len =	1584	nex =	1	
	Sngl	60820	59237	-	0
30	>2160132	/21783			
	len =	415	nex =	1	
	Sngl	78298	78712	+	0
35	>2160132	/21416			
	len =	698	nex =	1	
40	Sngl	79001	78304	-	0
	>2160132	/15957			
	len =	2326	nex =	10	
45	Term	88656	88489	-	0
	Intr	88915	88769	-	0
	Intr	89076	89020	-	0
	Intr	89255	89214	-	0
50	Intr	89496	89416	-	0
	Intr	89765	89687	-	0
	Intr	89957	89886	-	0
	Intr	90142	90042	-	0
	Intr	90343	90314	-	0
55	Init	90814	90417	-	0
	>2160132	/41375			
	len =	2353	nex =	10	
60					

					904
	Term	88656	88487	-	0
	Intr	88915	88769	-	0
	Intr	89076	89020	-	0
	Intr	89255	89214	-	0
5	Intr	89496	89416	-	0
	Intr	89765	89687	-	0
	Intr	89957	89886	-	0
	Intr	90142	90042	-	0
	Intr	90343	90314	-	0
10	Init	90839	90417	-	0
	>2160155	/17761			
15	len =	3312	nex =	7	
	Term	15073	14667	-	0
	Intr	15288	15145	-	0
	Intr	15853	15737	-	0
	Intr	16067	15930	-	0
20	Intr	16314	16190	-	0
	Intr	17071	16986	-	0
	Init	17685	17547	-	0
	>2160155	/6295			
25	len =	1644	nex =	2	
	Init	44387	45479	+	0
30	Term	45550	46030	+	0
	>2160155	/22067			
	len =	2477	nex =	7	
35	Term	52597	52279	-	0
	Intr	52839	52684	-	0
	Intr	53020	52925	-	0
	Intr	53302	53110	-	0
	Intr	54097	53469	-	0
40	Intr	54467	54189	-	0
	Init	54755	54540	-	0
	>2160155	/6319			
45	len =	3299	nex =	11	
	Init	60441	60770	+	0
	Intr	61340	61399	+	0
	Intr	61506	61619	+	0
50	Intr	61883	61948	+	0
	Intr	62027	62134	+	0
	Intr	62237	62320	+	0
	Intr	62639	62740	+	0
	Intr	62828	62941	+	0
55	Intr	63016	63096	+	0
	Intr	63191	63310	+	0
	Term	63481	63739	+	0
60	>2160155	/17081			

905

	len =	1390	nex =	2	
	Term	5724	5111	-	0
	Init	6498	6293	-	0
5	>2160155 /39525				
	len =	2602	nex =	6	
10	Init	7274	7423	+	0
	Intr	7512	7572	+	0
	Intr	7673	7725	+	0
	Intr	7845	7946	+	0
	Intr	8057	8546	+	0
15	Term	8659	9410	+	0
	>2160155 /6642				
	len =	823	nex =	2	
20	Init	76276	76526	+	0
	Term	76743	77098	+	0
	>2160155 /8575				
25	len =	851	nex =	2	
	Init	76277	76526	+	0
	Term	76743	77127	+	0
30	>2160155 /34772				
	len =	1361	nex =	3	
35	Init	7845	7946	+	0
	Intr	8057	8546	+	0
	Term	8659	9205	+	0
	>2160155 /23319				
40	len =	798	nex =	2	
	Term	86139	86067	-	0
	Init	86864	86267	-	0
45	>2182285 /21725				
	len =	2410	nex =	4	
50	Init	10500	10780	+	0
	Intr	11596	11657	+	0
	Intr	12371	12411	+	0
	Term	12536	12907	+	0
55	>2182285 /2118				
	len =	508	nex =	2	
	Init	33841	33970	+	0
60	Term	34088	34262	+	0

	>2182285	/25136			
5	len =	674	nex =	2	
	Init	36937	37065	+	0
	Term	37294	37610	+	0
10	>2182285	/108302			
	len =	610	nex =	2	
	Init	38504	38637	+	0
15	Term	38787	39106	+	0
	>2182285	/1264			
	len =	819	nex =	1	
20	Sngl	40950	40138	-	0
	>2182285	/27763			
25	len =	2303	nex =	5	
	Term	51059	50550	-	0
	Intr	51488	51406	-	0
	Intr	51733	51567	-	0
	Intr	52004	51818	-	0
30	Init	52852	52510	-	0
	>2182285	/13186			
35	len =	2050	nex =	0	
	>2182285	/27609			
	len =	957	nex =	4	
40	Term	97439	97181	-	0
	Intr	97605	97537	-	0
	Intr	97796	97693	-	0
	Init	98125	97940	-	0
45	>2182286	/37761			
	len =	1581	nex =	2	
50	Term	12027	10760	-	0
	Init	12340	12113	-	0
	>2182286	/34835			
55	len =	2290	nex =	6	
	Term	20929	20543	-	0
	Intr	21150	21014	-	0
	Intr	21441	21240	-	0
	Intr	21635	21537	-	0
60	Intr	22248	22162	-	0

				907	
	Init	22824	22577	-	0
	>2182286	/15161			
5	len =	336	nex =	1	
	Sngl	32955	32620	-	0
	>2182286	/3538			
10	len =	1090	nex =	5	
	Term	56957	56571	-	0
	Intr	57090	57034	-	0
15	Intr	57291	57192	-	0
	Intr	57436	57378	-	0
	Init	57657	57525	-	0
	>2182286	/31705			
20	len =	1400	nex =	1	
	Sngl	61850	60451	-	0
25	>2182287	/13008			
	len =	2683	nex =	13	
	Init	15455	15581	+	0
30	Intr	15687	15748	+	0
	Intr	15834	15911	+	0
	Intr	15991	16066	+	0
	Intr	16164	16234	+	0
	Intr	16347	16448	+	0
35	Intr	16539	16628	+	0
	Intr	16756	16830	+	0
	Intr	17067	17099	+	0
	Intr	17201	17272	+	0
	Intr	17463	17525	+	0
40	Intr	17658	17773	+	0
	Term	17832	18137	+	0
	>2182287	/14016			
45	len =	1166	nex =	1	
	Sngl	33340	33469	+	0
	>2182287	/35042			
50	len =	497	nex =	1	
	Sngl	33706	33927	+	0
55	>2182287	/20858			
	len =	1183	nex =	1	
	Sngl	50135	49403	-	0
60					

	>2182287	/2985			
	len =	1215	nex =	2	
5	Init	66989	67367	+	0
	Term	67700	68203	+	0
	>2182287	/20754			
10	len =	1074	nex =	2	
	Init	67134	67367	+	0
	Term	67700	68203	+	0
15	>2182287	/13385			
	len =	3085	nex =	14	
	Init	70516	70728	+	0
20	Intr	71073	71225	+	0
	Intr	71312	71410	+	0
	Intr	71569	71670	+	0
	Intr	71750	71893	+	0
	Intr	72009	72103	+	0
25	Intr	72210	72339	+	0
	Intr	72415	72519	+	0
	Intr	72624	72730	+	0
	Intr	72806	72857	+	0
	Intr	72941	73030	+	0
30	Intr	73120	73204	+	0
	Intr	73290	73354	+	0
	Term	73441	73600	+	0
	>2182287	/121535			
35	len =	153	nex =	1	
	Sngl	88279	88127	-	0
40	>2182287	/103034			
	len =	910	nex =	3	
	Term	97140	96927	-	0
45	Intr	97606	97230	-	0
	Init	97824	97721	-	0
	>2182289	/27197			
50	len =	2831	nex =	11	
	Term	14016	13844	-	0
	Intr	14182	14117	-	0
	Intr	14330	14282	-	0
55	Intr	14688	14621	-	0
	Intr	14820	14775	-	0
	Intr	15017	14924	-	0
	Intr	15178	15116	-	0
	Intr	15465	15343	-	0
60	Intr	15654	15621	-	0

					909
	Intr	16127	16011	-	0
	Init	16674	16500	-	0
	>2182289	/205500			
5	len =	1030	nex =	1	
	Sngl	52945	51919	-	0
10	>2182289	/31372			
	len =	1810	nex =	7	
	Term	58360	58263	-	0
15	Intr	58757	58601	-	0
	Intr	59020	58889	-	0
	Intr	59232	59120	-	0
	Intr	59487	59328	-	0
	Intr	59636	59613	-	0
20	Init	60069	59921	-	0
	>2182289	/38858			
	len =	1690	nex =	3	
25	Term	85015	84307	-	0
	Intr	85533	85226	-	0
	Init	85988	85678	-	0
30	>2191126	/28640			
	len =	1810	nex =	3	
	Init	105687	105961	+	0
35	Intr	106179	106465	+	0
	Term	106570	107490	+	0
	>2191126	/1204			
40	len =	1690	nex =	1	
	Sngl	110532	112213	+	0
	>2191126	/41187			
45	len =	2551	nex =	7	
	Term	115853	115628	-	0
	Intr	116345	116037	-	0
50	Intr	116498	116418	-	0
	Intr	116825	116751	-	0
	Intr	116990	116904	-	0
	Intr	117192	117090	-	0
	Init	118178	117485	-	0
55	>2191126	/21195			
	len =	2566	nex =	7	
60	Term	115853	115637	-	0

910

	Intr	116345	116037	-	0
	Intr	116498	116418	-	0
	Intr	116825	116751	-	0
	Intr	116990	116904	-	0
5	Intr	117192	117090	-	0
	Init	118202	117485	-	0
	>2191126		/19141		
10	len =	3438	nex =	14	
	Term	25095	24742	-	0
	Intr	25245	25180	-	0
	Intr	25409	25338	-	0
15	Intr	25625	25512	-	0
	Intr	25812	25720	-	0
	Intr	25961	25899	-	0
	Intr	26152	26042	-	0
	Intr	26360	26247	-	0
20	Intr	26604	26506	-	0
	Intr	26756	26691	-	0
	Intr	26948	26853	-	0
	Intr	27119	27035	-	0
	Intr	27350	27203	-	0
25	Init	28179	28046	-	0
	>2191126		/22919		
	len =	1497	nex =	4	
30	Init	28448	28746	+	0
	Intr	29035	29235	+	0
	Intr	29321	29463	+	0
	Term	29617	29944	+	0
35	>2191126		/117191		
	len =	253	nex =	1	
40	Sngl	66070	66322	+	0
	>2191126		/7653		
	len =	1991	nex =	6	
45	Term	5264	4896	-	0
	Intr	5520	5337	-	0
	Intr	5933	5601	-	0
	Intr	6260	6123	-	0
50	Intr	6458	6329	-	0
	Init	6886	6590	-	0
	>2191126		/41270		
55	len =	1008	nex =	4	
	Term	79378	79212	-	0
	Intr	79532	79461	-	0
	Intr	79730	79623	-	0
60	Init	79919	79819	-	0

>2191126 /94836

5	len =	1131	nex =	5	
	Term	79378	79190	-	0
	Intr	79532	79461	-	0
	Intr	79730	79623	-	0
	Intr	79919	79819	-	0
10	Init	80320	80142	-	0

>2191126 /12604

15	len =	1177	nex =	5	
	Term	79378	79212	-	0
	Intr	79532	79461	-	0
	Intr	79730	79623	-	0
	Intr	79919	79819	-	0
20	Init	80388	80142	-	0

>2191126 /16533

25	len =	2830	nex =	12	
	Init	90009	90076	+	0
	Intr	90175	90222	+	0
	Intr	90315	90410	+	0
	Intr	90483	90563	+	0
30	Intr	90659	90700	+	0
	Intr	90784	90947	+	0
	Intr	91030	91078	+	0
	Intr	91166	91214	+	0
	Intr	91306	91445	+	0
35	Intr	91538	91615	+	0
	Intr	91709	91834	+	0
	Term	91909	92323	+	0

>2191157 /5457

40	len =	688	nex =	2	
	Term	110545	110202	-	0
45	Init	110889	110723	-	0

>2191157 /39714

	len =	520	nex =	1	
50	Sngl	24526	25045	+	0

>2191157 /37336

55	len =	1558	nex =	2	
	Init	26629	26769	+	0
	Term	27064	27170	+	0

>2191157 /17739

60					
----	--	--	--	--	--

912

	len =	2326	nex =	11	
	Term	1098	904	-	0
	Intr	1303	1201	-	0
5	Intr	1501	1418	-	0
	Intr	1698	1603	-	0
	Intr	1848	1798	-	0
	Intr	2076	1936	-	0
	Intr	2220	2164	-	0
10	Intr	2391	2317	-	0
	Intr	2739	2467	-	0
	Intr	2894	2835	-	0
	Init	3094	3002	-	0
15	>2191157 /21258				
	len =	2364	nex =	9	
	Init	35554	35767	+	0
20	Intr	35854	35917	+	0
	Intr	36017	36231	+	0
	Intr	36362	36538	+	0
	Intr	36622	36696	+	0
	Intr	36794	36895	+	0
25	Intr	37265	37376	+	0
	Intr	37474	37620	+	0
	Term	37753	37793	+	0
30	>2191157 /42174				
	len =	540	nex =	1	
	Sngl	59287	59826	+	0
35	>2191157 /27625				
	len =	732	nex =	2	
	Init	80900	81166	+	0
40	Term	81274	81631	+	0
	>2191157 /41361				
45	len =	2136	nex =	2	
	Init	83526	83731	+	0
	Term	83861	84187	+	0
50	>2191157 /32265				
	len =	2008	nex =	2	
	Init	83526	83731	+	0
	Term	83861	84181	+	0
55	>2191157 /2495				
	len =	2795	nex =	5	
60	Init	92543	92875	+	0

				913	
	Intr	93634	93776	+	0
	Intr	94054	94077	+	0
	Intr	94512	94714	+	0
	Term	94965	95337	+	0
5	>2191181	/38304			
	len =	2070	nex =	4	
10	Init	1742	2050	+	0
	Intr	2468	2686	+	0
	Intr	2758	2844	+	0
	Term	3193	3219	+	0
15	>2191181	/23239			
	len =	988	nex =	3	
	Term	4337	3802	-	0
20	Intr	4497	4418	-	0
	Init	4789	4601	-	0
	>2191181	/30935			
25	len =	1455	nex =	0	
	>2213606	/6503			
	len =	1974	nex =	4	
30	Init	15815	16171	+	0
	Intr	16373	16842	+	0
	Intr	16925	17188	+	0
	Term	17281	17788	+	0
35	>2213606	/10990			
	len =	413	nex =	1	
40	Sngl	18252	17840	-	0
	>2213606	/38093			
	len =	490	nex =	1	
45	Sngl	27032	27514	+	0
	>2213606	/23231			
50	len =	700	nex =	1	
	Sngl	45292	44593	-	0
	>2213606	/31944			
55	len =	559	nex =	1	
	Sngl	49930	49372	-	0
60	>2244747	/16846			

	len =	2017	nex =	2	
5	Init	12786	13565	+	0
	Term	13854	14802	+	0
	>2244747 /38987				
10	len =	134	nex =	1	
	Sngl	14762	14895	+	0
	>2244747 /17977				
15	len =	610	nex =	1	
	Sngl	16599	15997	-	0
20	>2244747 /19172				
	len =	610	nex =	1	
	Sngl	16614	16009	-	0
25	>2244747 /30129				
	len =	813	nex =	1	
30	Sngl	176792	177114	+	0
	>2244747 /195				
	len =	805	nex =	1	
35	Sngl	176309	177113	+	0
	>2244747 /101734				
40	len =	340	nex =	1	
	Sngl	198899	199238	+	0
	>2244747 /126389				
45	len =	1776	nex =	8	
	Init	48741	48903	+	0
	Intr	48995	49057	+	0
	Intr	49141	49207	+	0
50	Intr	49296	49396	+	0
	Intr	49486	49530	+	0
	Intr	49614	49895	+	0
	Intr	49983	50085	+	0
	Term	50189	50516	+	0
55	>2244747 /25991				
	len =	1850	nex =	8	
60	Init	48741	48903	+	0

				915
	Intr	48995	49057	+ 0
	Intr	49141	49207	+ 0
	Intr	49296	49396	+ 0
	Intr	49486	49530	+ 0
5	Intr	49614	49895	+ 0
	Intr	49983	50085	+ 0
	Term	50189	50590	+ 0
	>2244747 /99093			
10	len =	430	nex =	3
	Init	48743	48903	+ 0
	Intr	48995	49057	+ 0
15	Term	49141	49172	+ 0
	>2244747 /7346			
	len =	550	nex =	1
20	Sngl	51305	50761	- 0
	>2244747 /13520			
25	len =	522	nex =	1
	Sngl	53660	53139	- 0
	>2244747 /18697			
30	len =	817	nex =	2
	Term	56326	55871	- 0
	Init	56687	56413	- 0
35	>2244747 /35186			
	len =	1525	nex =	5
40	Term	56326	55870	- 0
	Intr	56685	56413	- 0
	Intr	56884	56777	- 0
	Intr	57220	56989	- 0
	Init	57394	57303	- 0
45	>2244747 /39975			
	len =	2277	nex =	7
50	Term	56326	55859	- 0
	Intr	56685	56413	- 0
	Intr	56884	56777	- 0
	Intr	57220	56989	- 0
	Intr	57530	57303	- 0
55	Intr	57816	57621	- 0
	Init	58135	57936	- 0
	>2244747 /108308			
60	len =	2306	nex =	6

	Term	58896	58494	-	0
	Intr	59256	58984	-	0
	Intr	59446	59412	-	0
5	Intr	59994	59535	-	0
	Intr	60270	60075	-	0
	Init	60799	60608	-	0
>2244747 /34967					
10	len =	1692	nex =	3	
	Init	78644	78978	+	0
	Intr	79811	79967	+	0
15	Term	80055	80335	+	0
>2244747 /29662					
	len =	2324	nex =	4	
20	Term	6181	5707	-	0
	Intr	6376	6275	-	0
	Intr	6858	6468	-	0
	Init	8030	7268	-	0
25	>2244747 /10852				
	len =	948	nex =	3	
30	Term	95484	95087	-	0
	Intr	95756	95563	-	0
	Init	96034	95845	-	0
>2244747 /33554					
35	len =	1225	nex =	3	
	Term	95484	94981	-	0
	Intr	95756	95563	-	0
40	Init	96205	95845	-	0
>2244788 /33860					
	len =	894	nex =	2	
45	Init	119066	119340	+	0
	Term	119433	119959	+	0
>2244788 /4232					
50	len =	1570	nex =	3	
	Term	11837	11610	-	0
	Intr	12997	12874	-	0
55	Init	13171	13086	-	0
>2244788 /20129					
60	len =	1736	nex =	4	

```

                                     917
      Init  134496  134633      +      0
      Intr  134785  134908      +      0
      Intr  135250  135306      +      0
      Term  135918  136231      +      0
5
>2244788      /4905
      len =    1532    nex =      4

10      Init  134547  134633      +      0
      Intr  134785  134908      +      0
      Intr  135250  135306      +      0
      Term  135918  136078      +      0

15 >2244788      /18255
      len =    1917    nex =      4

      Term  11837   11553      -      0
20      Intr  12997   12874      -      0
      Intr  13171   13086      -      0
      Init  13469   13401      -      0

      >2244788      /42223
25      len =    1270    nex =      2

      Init  141770   141970      +      0
      Term  142713   143034      +      0
30 >2244788      /21908
      len =     865    nex =      2

35      Term  172609   172540      -      0
      Init  173404   172806      -      0

      >2244788      /95834
40      len =     932    nex =      4

      Init  176283   176507      +      0
      Intr  176602   176703      +      0
      Intr  176785   176939      +      0
45      Term  176951   177214      +      0

      >2244788      /31495
      len =    1150    nex =      5
50      Init  177820   177887      +      0
      Intr  178110   178208      +      0
      Intr  178295   178347      +      0
      Intr  178445   178518      +      0
55      Term  178797   178969      +      0

      >2244788      /40073
      len =    1761    nex =      5
60

```

918

	Term	182960	182681	-	0
	Intr	183144	183074	-	0
	Intr	183352	183228	-	0
	Intr	183544	183430	-	0
5	Init	184441	183731	-	0
	>2244788		/2738		
10	len =	1855	nex =	7	
	Term	182960	182701	-	0
	Intr	183144	183074	-	0
	Intr	183352	183228	-	0
	Intr	183544	183430	-	0
15	Intr	183825	183731	-	0
	Intr	184012	183901	-	0
	Init	184555	184343	-	0
20	>2244788		/18153		
	len =	1337	nex =	4	
	Term	744	549	-	0
	Intr	903	829	-	0
25	Intr	1232	1053	-	0
	Init	1885	1804	-	0
	>2244788		/16319		
30	len =	1732	nex =	7	
	Term	188526	188214	-	0
	Intr	188710	188640	-	0
	Intr	188914	188790	-	0
35	Intr	189112	188998	-	0
	Intr	189340	189246	-	0
	Intr	189532	189421	-	0
	Init	189945	189850	-	0
40	>2244788		/34477		
	len =	790	nex =	4	
	Term	26188	26035	-	0
45	Intr	26496	26276	-	0
	Intr	26702	26590	-	0
	Init	26822	26779	-	0
50	>2244788		/37809		
	len =	2215	nex =	10	
	Term	29960	29503	-	0
	Intr	30139	30054	-	0
55	Intr	30309	30235	-	0
	Intr	30490	30388	-	0
	Intr	30687	30606	-	0
	Intr	30881	30790	-	0
	Intr	31057	30969	-	0
60	Intr	31236	31156	-	0

919

	Intr	31450	31336	-	0
	Init	31717	31579	-	0
	>2244788	/9870			
5	len =	1700	nex =	6	
	Term	45280	45046	-	0
	Intr	45431	45380	-	0
10	Intr	45545	45518	-	0
	Intr	46149	46080	-	0
	Intr	46413	46313	-	0
	Init	46745	46519	-	0
15	>2244788	/40736			
	len =	1713	nex =	5	
	Init	57948	58133	+	0
20	Intr	58560	58765	+	0
	Intr	58850	58930	+	0
	Intr	59012	59174	+	0
	Term	59262	59660	+	0
25	>2244788	/1718			
	len =	1844	nex =	5	
	Term	60276	59985	-	0
30	Intr	60467	60369	-	0
	Intr	60644	60555	-	0
	Intr	60856	60742	-	0
	Init	61828	61672	-	0
35	>2244788	/94503			
	len =	1930	nex =	5	
	Term	60276	59949	-	0
40	Intr	60467	60369	-	0
	Intr	60644	60555	-	0
	Intr	60856	60742	-	0
	Init	61875	61672	-	0
45	>2244788	/28978			
	len =	921	nex =	1	
	Sngl	63706	62786	-	0
50	>2244788	/36844			
	len =	1309	nex =	1	
55	Sngl	78815	80123	+	0
	>2244788	/42933			
60	len =	2960	nex =	6	

920

	Init	92232	92765	+	0
	Intr	92959	93121	+	0
	Intr	93567	93743	+	0
	Intr	93831	93914	+	0
5	Intr	94438	94519	+	0
	Term	94602	95191	+	0
>2244829 /38042					
10	len =	2717	nex =	9	
	Init	103735	104049	+	0
	Intr	104329	104423	+	0
	Intr	104545	104609	+	0
15	Intr	104833	104876	+	0
	Intr	105212	105295	+	0
	Intr	105486	105639	+	0
	Intr	105738	105920	+	0
	Intr	106013	106069	+	0
20	Term	106159	106451	+	0
>2244829 /293					
	len =	315	nex =	1	
25	Sngl	114012	113698	-	0
>2244829 /40074					
30	len =	1498	nex =	2	
	Term	115095	113973	-	0
	Init	115470	115294	-	0
35	>2244829 /38411				
	len =	1796	nex =	2	
	Term	115095	113698	-	0
40	Init	115493	115294	-	0
>2244829 /10518					
	len =	2190	nex =	8	
45	Init	116378	116531	+	0
	Intr	116787	116872	+	0
	Intr	116953	117024	+	0
	Intr	117143	117180	+	0
50	Intr	117526	117569	+	0
	Intr	117791	117837	+	0
	Intr	117992	118166	+	0
	Term	118269	118567	+	0
55	>2244829 /29288				
	len =	492	nex =	1	
	Sngl	131227	130736	-	0
60					

	>2244829	/24175			
	len =	332	nex =	1	
5	Sngl	136899	137230	+	0
	>2244829	/17179			
	len =	450	nex =	1	
10	Sngl	136899	137332	+	0
	>2244829	/99523			
	len =	346	nex =	1	
15	Sngl	136900	137245	+	0
	>2244829	/37184			
20	len =	624	nex =	0	
	>2244829	/126602			
	len =	654	nex =	1	
25	Sngl	136900	137553	+	0
	>2244829	/15384			
30	len =	627	nex =	1	
	Sngl	136904	137530	+	0
	>2244829	/26797			
35	len =	628	nex =	1	
	Sngl	136904	137531	+	0
40	>2244829	/36129			
	len =	739	nex =	1	
45	Sngl	199828	200566	+	0
	>2244829	/24266			
	len =	1908	nex =	3	
50	Init	65354	65621	+	0
	Intr	65713	65836	+	0
	Term	66807	67261	+	0
	>2244829	/31856			
	len =	897	nex =	3	
	Init	70117	70500	+	0
60	Intr	70585	70611	+	0

					922
	Term	70696	71013	+	0
	>2244829	/30327			
5	len =	711	nex =	1	
	Sngl	82258	82968	+	0
	>2244829	/33166			
10	len =	650	nex =	1	
	Sngl	82303	82952	+	0
	>2244829	/42848			
15	len =	2473	nex =	9	
	Term	83367	83062	-	0
20	Intr	83556	83476	-	0
	Intr	83703	83644	-	0
	Intr	83890	83811	-	0
	Intr	84071	84020	-	0
	Intr	84306	84169	-	0
25	Intr	84661	84398	-	0
	Intr	84799	84742	-	0
	Init	84996	84887	-	0
	>2244829	/22861			
30	len =	611	nex =	1	
	Sngl	85902	86512	+	0
	>2244829	/25333			
35	len =	2115	nex =	3	
	Term	87340	86629	-	0
40	Intr	87618	87443	-	0
	Init	88743	87767	-	0
	>2244829	/117350			
45	len =	1760	nex =	8	
	Term	93545	93422	-	0
	Intr	93819	93710	-	0
	Intr	93998	93936	-	0
50	Intr	94168	94094	-	0
	Intr	94368	94276	-	0
	Intr	94573	94469	-	0
	Intr	94861	94740	-	0
	Init	95181	94950	-	0
55	>2244870	/2163			
	len =	1517	nex =	1	
60	Sngl	13507	15023	+	0

	>2244870	/15641			
5	len =	1853	nex =	7	
	Init	2352	2569	+	0
	Intr	2668	2781	+	0
	Intr	2862	2957	+	0
	Intr	3057	3099	+	0
10	Intr	3174	3326	+	0
	Intr	3408	3476	+	0
	Term	3843	4204	+	0
	>2244870	/35290			
15	len =	1090	nex =	2	
	Term	33366	33045	-	0
20	Init	34113	33943	-	0
	>2244870	/18642			
	len =	867	nex =	2	
25	Term	4431	4071	-	0
	Init	4937	4513	-	0
	>2244870	/30852			
30	len =	513	nex =	1	
	Sngl	70945	70433	-	0
	>2244870	/36205			
35	len =	1210	nex =	1	
	Sngl	71644	70435	-	0
40	>2244870	/30929			
	len =	867	nex =	1	
	Sngl	84563	85414	+	0
45	>2244901	/32219			
	len =	644	nex =	1	
50	Sngl	100297	100940	+	0
	>2244901	/101301			
55	len =	1235	nex =	2	
	Init	12251	12597	+	0
	Term	13371	13485	+	0
60	>2244901	/15334			

924

	len =	2089	nex =	4	
	Init	12251	12597	+	0
	Intr	13371	13484	+	0
5	Intr	13678	13835	+	0
	Term	13944	14339	+	0
	>2244901 /14485				
10	len =	1048	nex =	2	
	Term	136645	136202	-	0
	Init	137249	136976	-	0
15	>2244901 /8916				
	len =	761	nex =	2	
	Init	146636	146871	+	0
20	Term	146912	147396	+	0
	>2244901 /22637				
	len =	1930	nex =	7	
25	Init	150934	151112	+	0
	Intr	151807	151845	+	0
	Intr	151938	151991	+	0
	Intr	152091	152144	+	0
30	Intr	152269	152322	+	0
	Intr	152417	152488	+	0
	Term	152622	152862	+	0
	>2244901 /5455				
35	len =	550	nex =	1	
	Sngl	153514	154059	+	0
40	>2244901 /25390				
	len =	1731	nex =	3	
	Term	156239	156216	-	0
45	Intr	156385	156332	-	0
	Init	157099	156997	-	0
	>2244901 /39757				
50	len =	1489	nex =	5	
	Term	164193	163773	-	0
	Intr	164487	164293	-	0
	Intr	164750	164603	-	0
55	Intr	164938	164832	-	0
	Init	165261	165017	-	0
	>2244901 /113295				
60	len =	250	nex =	1	

925

	Sngl	165261	165021	-	0
	>2244901	/43007			
5	len =	3418	nex =	9	
	Init	181307	182180	+	0
	Intr	182482	182558	+	0
10	Intr	182639	182732	+	0
	Intr	182817	182915	+	0
	Intr	183212	183301	+	0
	Intr	183400	183519	+	0
	Intr	183767	183870	+	0
15	Intr	184163	184235	+	0
	Term	184397	184724	+	0
	>2244901	/8381			
20	len =	928	nex =	2	
	Init	197128	197392	+	0
	Term	197699	198055	+	0
25	>2244901	/35383			
	len =	1690	nex =	1	
	Sngl	23032	21343	-	0
30	>2244901	/12451			
	len =	2050	nex =	4	
35	Init	29261	29459	+	0
	Intr	29681	29785	+	0
	Intr	29969	30397	+	0
	Term	30959	31303	+	0
40	>2244901	/8234			
	len =	855	nex =	4	
	Term	33518	33296	-	0
45	Intr	33802	33633	-	0
	Intr	34017	33880	-	0
	Init	34150	34103	-	0
	>2244901	/33073			
50	len =	3028	nex =	2	
	Init	4164	4631	+	0
	Term	6071	7191	+	0
55	>2244901	/307			
	len =	1838	nex =	8	
60	Init	44565	44888	+	0

					926
	Intr	44976	45044	+	0
	Intr	45145	45198	+	0
	Intr	45288	45327	+	0
	Intr	45414	45512	+	0
5	Intr	45595	45819	+	0
	Intr	45902	46023	+	0
	Term	46120	46402	+	0
	>2244901	/19122			
10	len =	1766	nex =	8	
	Init	44638	44888	+	0
	Intr	44976	45044	+	0
15	Intr	45145	45198	+	0
	Intr	45288	45327	+	0
	Intr	45414	45512	+	0
	Intr	45595	45819	+	0
	Intr	45902	46023	+	0
20	Term	46120	46403	+	0
	>2244901	/37345			
	len =	1379	nex =	3	
25	Init	55027	55308	+	0
	Intr	55387	55671	+	0
	Term	55759	56179	+	0
30	>2244901	/26019			
	len =	1750	nex =	3	
	Init	77747	78039	+	0
35	Intr	78780	78906	+	0
	Term	79065	79492	+	0
	>2244901	/933			
40	len =	1415	nex =	2	
	Init	86075	86413	+	0
	Term	86998	87489	+	0
45	>2244950	/12629			
	len =	3346	nex =	10	
	Term	100982	100625	-	0
50	Intr	101466	101106	-	0
	Intr	101718	101591	-	0
	Intr	102002	101874	-	0
	Intr	102439	102360	-	0
	Intr	102690	102527	-	0
55	Intr	102958	102773	-	0
	Intr	103205	103074	-	0
	Intr	103432	103291	-	0
	Init	103970	103568	-	0
60	>2244950	/40414			

	len =	2150	nex =	6	
	Term	109338	109067	-	0
5	Intr	109551	109489	-	0
	Intr	109708	109646	-	0
	Intr	109850	109803	-	0
	Intr	110001	109939	-	0
	Init	111043	110961	-	0
10	>2244950 /30227				
	len =	2050	nex =	6	
15	Term	109338	109187	-	0
	Intr	109551	109489	-	0
	Intr	109708	109646	-	0
	Intr	109850	109803	-	0
	Intr	110001	109939	-	0
20	Init	111043	110961	-	0
	>2244950 /5714				
	len =	1403	nex =	7	
25	Init	124186	124326	+	0
	Intr	124418	124469	+	0
	Intr	124596	124670	+	0
	Intr	124766	124794	+	0
30	Intr	124968	125001	+	0
	Intr	125082	125152	+	0
	Term	125251	125588	+	0
	>2244950 /33513				
35	len =	1593	nex =	4	
	Init	138127	138644	+	0
	Intr	138739	138858	+	0
40	Intr	138934	139180	+	0
	Term	139256	139719	+	0
	>2244950 /19028				
45	len =	638	nex =	2	
	Init	139024	139180	+	0
	Term	139256	139661	+	0
50	>2244950 /21894				
	len =	1030	nex =	1	
	Sngl	146832	145803	-	0
55	>2244950 /7605				
	len =	814	nex =	2	
60	Term	167332	166714	-	0

				928	
	Init	167527	167451	-	0
	>2244950	/3176			
5	len =	1423	nex =	2	
	Term	167332	166764	-	0
	Init	167934	167451	-	0
10	>2244950	/41791			
	len =	1479	nex =	3	
	Term	167332	166712	-	0
15	Intr	167934	167451	-	0
	Init	168190	168116	-	0
	>2244950	/12256			
20	len =	1716	nex =	4	
	Term	169269	169015	-	0
	Intr	169606	169448	-	0
	Intr	170335	170260	-	0
25	Init	170730	170607	-	0
	>2244950	/6723			
	len =	1536	nex =	4	
30	Init	171676	171958	+	0
	Intr	172224	172415	+	0
	Intr	172496	172661	+	0
	Term	172740	173211	+	0
35	>2244950	/124835			
	len =	978	nex =	1	
40	Sngl	18831	19808	+	0
	>2244950	/40793			
	len =	1247	nex =	3	
45	Term	193189	192906	-	0
	Intr	193587	193266	-	0
	Init	194152	193673	-	0
50	>2244950	/2803			
	len =	1824	nex =	3	
	Init	2896	3184	+	0
55	Intr	3571	3676	+	0
	Term	4403	4719	+	0
	>2244950	/9209			
60	len =	573	nex =	1	

929

	Sngl	31137	30565	-	0
	>2244950	/29655			
5	len =	682	nex =	1	
	Sngl	34486	35167	+	0
10	>2244950	/40913			
	len =	2079	nex =	7	
	Init	4949	5128	+	0
15	Intr	5254	5419	+	0
	Intr	5498	5550	+	0
	Intr	5911	5973	+	0
	Intr	6366	6416	+	0
	Intr	6516	6630	+	0
20	Term	6687	7027	+	0
	>2244950	/18234			
	len =	1950	nex =	6	
25	Init	61059	61335	+	0
	Intr	61420	61550	+	0
	Intr	61714	61791	+	0
	Intr	61882	61926	+	0
30	Intr	62016	62060	+	0
	Term	62293	62389	+	0
	>2244950	/32203			
35	len =	1510	nex =	6	
	Init	7376	7454	+	0
	Intr	7542	7577	+	0
	Intr	7707	7844	+	0
40	Intr	7939	8012	+	0
	Intr	8418	8486	+	0
	Term	8556	8884	+	0
	>2244950	/31782			
45	len =	1211	nex =	1	
	Sngl	84183	82973	-	0
50	>2244950	/17019			
	len =	2897	nex =	2	
	Term	84672	82981	-	0
55	Init	85877	85235	-	0
	>2244950	/109560			
	len =	397	nex =	1	
60					

930

	Sngl	95604	96000	+	0
	>2244991	/7101			
5	len =	1300	nex =	5	
	Term	99473	99160	-	0
	Intr	99674	99597	-	0
	Intr	99851	99788	-	0
10	Intr	100015	99939	-	0
	Init	100216	100170	-	0
	>2244991	/14136			
15	len =	1251	nex =	1	
	Sngl	133001	131751	-	0
	>2244991	/24611			
20	len =	1275	nex =	6	
	Init	144816	144916	+	0
	Intr	144996	145065	+	0
25	Intr	145153	145209	+	0
	Intr	145299	145360	+	0
	Intr	145408	145507	+	0
	Term	145593	145964	+	0
30	>2244991	/5546			
	len =	1163	nex =	3	
	Term	157187	156808	-	0
35	Intr	157430	157305	-	0
	Init	157970	157545	-	0
	>2244991	/8212			
40	len =	1254	nex =	0	
	>2244991	/40778			
	len =	879	nex =	3	
45	Init	163368	163492	+	0
	Intr	163658	163757	+	0
	Term	163863	164240	+	0
50	>2244991	/23771			
	len =	1377	nex =	4	
	Term	164902	164507	-	0
55	Intr	165186	164989	-	0
	Intr	165666	165500	-	0
	Init	165883	165813	-	0
	>2244991	/16525			
60					

931

	len =	810	nex =	2	
	Init	172277	172503	+	0
	Term	172604	173086	+	0
5	>2244991 /22084				
	len =	1450	nex =	2	
10	Init	177203	177333	+	0
	Term	177407	177827	+	0
	>2244991 /157870				
15	len =	342	nex =	1	
	Sngl	17882	17541	-	0
	>2244991 /5686				
20	len =	2453	nex =	10	
	Term	194540	194396	-	0
	Intr	194759	194680	-	0
25	Intr	194888	194843	-	0
	Intr	195027	194971	-	0
	Intr	195163	195105	-	0
	Intr	195344	195244	-	0
	Intr	195623	195502	-	0
30	Intr	195980	195929	-	0
	Intr	196138	196058	-	0
	Init	196848	196213	-	0
	>2244991 /2505				
35	len =	623	nex =	1	
	Sngl	27093	26471	-	0
40	>2244991 /7632				
	len =	1210	nex =	3	
	Term	36794	36385	-	0
45	Intr	37205	37073	-	0
	Init	37590	37308	-	0
	>2244991 /30471				
50	len =	1883	nex =	6	
	Term	39363	38946	-	0
	Intr	39486	39437	-	0
	Intr	39651	39570	-	0
55	Intr	39806	39736	-	0
	Intr	40168	40098	-	0
	Init	40371	40292	-	0
	>2244991 /17535				
60					

932

	len =	585	nex =	1	
	Sngl	43288	43872	+	0
5	>2244991	/17553			
	len =	628	nex =	2	
	Init	44575	44786	+	0
10	Term	44876	45202	+	0
	>2244991	/16090			
	len =	634	nex =	2	
15	Init	44583	44786	+	0
	Term	44876	45216	+	0
	>2244991	/31946			
20	len =	562	nex =	1	
	Sngl	66524	65963	-	0
25	>2244991	/6580			
	len =	509	nex =	1	
	Sngl	70265	69757	-	0
30	>2244991	/17851			
	len =	1752	nex =	5	
35	Term	71484	71210	-	0
	Intr	71754	71636	-	0
	Intr	71898	71846	-	0
	Intr	72484	72429	-	0
	Init	72626	72579	-	0
40	>2244991	/92054			
	len =	587	nex =	1	
45	Sngl	8564	9150	+	0
	>2245031	/92144			
	len =	444	nex =	1	
50	Sngl	125198	125641	+	0
	>2245031	/30087			
55	len =	822	nex =	1	
	Sngl	125198	126019	+	0
60	>2245031	/118011			

	len =	355	nex =	1	
	Sngl	125287	125641	+	0
5	>2245031	/91870			
	len =	1970	nex =	4	
	Init	144106	144256	+	0
10	Intr	144641	144768	+	0
	Intr	145143	145253	+	0
	Term	145583	146075	+	0
	>2245031	/36017			
15	len =	3647	nex =	8	
	Term	154141	153926	-	0
	Intr	155021	154948	-	0
20	Intr	155252	155139	-	0
	Intr	155661	155584	-	0
	Intr	155955	155829	-	0
	Intr	156204	156149	-	0
	Intr	156561	156358	-	0
25	Init	157572	157241	-	0
	>2245031	/7834			
	len =	3010	nex =	12	
30	Init	157780	157908	+	0
	Intr	157993	158125	+	0
	Intr	158517	158604	+	0
	Intr	158708	158784	+	0
35	Intr	159068	159107	+	0
	Intr	159412	159497	+	0
	Intr	159590	159671	+	0
	Intr	159798	159854	+	0
	Intr	159938	159976	+	0
40	Intr	160067	160137	+	0
	Intr	160354	160407	+	0
	Term	160554	160780	+	0
	>2245031	/114540			
45	len =	3018	nex =	11	
	Init	157780	157908	+	0
	Intr	157993	158125	+	0
50	Intr	158517	158604	+	0
	Intr	158708	158784	+	0
	Intr	159068	159497	+	0
	Intr	159590	159671	+	0
	Intr	159798	159854	+	0
55	Intr	159938	159976	+	0
	Intr	160067	160137	+	0
	Intr	160354	160407	+	0
	Term	160554	160797	+	0
60	>2245031	/110681			

	len =	466	nex =	2	
	Init	172709	172801	+	0
5	Term	172906	173174	+	0
	>2245031	/142850			
	len =	610	nex =	1	
10	Sngl	173847	173242	-	0
	>2245031	/42533			
15	len =	1533	nex =	4	
	Init	17415	17660	+	0
	Intr	17764	18062	+	0
	Intr	18331	18410	+	0
20	Term	18499	18947	+	0
	>2245031	/36882			
25	len =	2299	nex =	5	
	Term	173963	173241	-	0
	Intr	174262	174007	-	0
	Intr	174516	174406	-	0
	Intr	174824	174614	-	0
30	Init	175539	174923	-	0
	>2245031	/14613			
35	len =	673	nex =	1	
	Sngl	20501	19829	-	0
	>2245031	/831			
40	len =	850	nex =	3	
	Init	39954	40111	+	0
	Intr	40198	40248	+	0
	Term	40330	40796	+	0
45	>2245031	/14223			
	len =	638	nex =	1	
50	Sngl	43095	43370	+	0
	>2245031	/35772			
	len =	1663	nex =	1	
55	Sngl	48986	49948	+	0
	>2245073	/158661			
60	len =	739	nex =	1	

935

	Sngl	102245	101507	-	0
	>2245073 /34167				
5	len =	1019	nex =	3	
	Init	104868	105196	+	0
	Intr	105282	105361	+	0
10	Term	105463	105866	+	0
	>2245073 /36603				
	len =	4481	nex =	11	
15	Term	6893	6584	-	0
	Intr	7287	7083	-	0
	Intr	7700	7618	-	0
	Intr	8129	7990	-	0
20	Intr	8424	8266	-	0
	Intr	9480	8479	-	0
	Intr	9839	9542	-	0
	Intr	10132	9928	-	0
	Intr	10433	10351	-	0
25	Intr	10748	10609	-	0
	Init	11064	10945	-	0
	>2245073 /37223				
30	len =	4483	nex =	11	
	Term	6893	6584	-	0
	Intr	7287	7083	-	0
	Intr	7700	7618	-	0
35	Intr	8129	7990	-	0
	Intr	8424	8266	-	0
	Intr	9480	8479	-	0
	Intr	9839	9542	-	0
	Intr	10132	9928	-	0
40	Intr	10433	10351	-	0
	Intr	10748	10609	-	0
	Init	11066	10945	-	0
	>2245073 /6042				
45	len =	959	nex =	1	
	Sngl	124096	125054	+	0
50	>2245073 /35156				
	len =	2133	nex =	7	
	Init	136139	136418	+	0
55	Intr	136654	136948	+	0
	Intr	137036	137101	+	0
	Intr	137200	137329	+	0
	Intr	137421	137579	+	0
	Intr	137703	137753	+	0
60	Term	137855	138271	+	0

```

>2245073      /154342
5      len =      111      nex =      1
      Sngl 140364 140254      -      0

>2245073      /3258
10     len =      2050     nex =      8
      Term 138586 138326      -      0
      Intr 138787 138684      -      0
      Intr 139039 138884      -      0
15     Intr 139188 139117      -      0
      Intr 139338 139291      -      0
      Intr 139469 139422      -      0
      Intr 139680 139608      -      0
      Init 140370 140183      -      0
20
>2245073      /2161
      len =      1690     nex =      5
25     Init 145051 145144      +      0
      Intr 145227 145544      +      0
      Intr 145712 145798      +      0
      Intr 145888 146021      +      0
      Term 146416 146733      +      0
30
>2245073      /17120
      len =      464      nex =      2
35     Init 145081 145144      +      0
      Term 145227 145544      +      0

>2245073      /29150
40     len =      1072     nex =      3
      Init 168520 168924      +      0
      Intr 169023 169160      +      0
      Term 169230 169591      +      0
45
>2245073      /23025
      len =      2715     nex =      8
50     Init 181224 181382      +      0
      Intr 181935 181992      +      0
      Intr 182407 182489      +      0
      Intr 182789 183061      +      0
      Intr 183152 183204      +      0
55     Intr 183325 183405      +      0
      Intr 183502 183614      +      0
      Term 183704 183938      +      0

>2245073      /19505
60

```

937

	len =	2035	nex =	4	
	Init	189969	190426	+	0
	Intr	190764	190988	+	0
5	Intr	191116	191225	+	0
	Term	191315	191480	+	0
	>2245073 /31781				
10	len =	1939	nex =	4	
	Init	190050	190426	+	0
	Intr	190764	190988	+	0
	Intr	191116	191225	+	0
15	Term	191315	191480	+	0
	>2245073 /36521				
20	len =	730	nex =	1	
	Sngl	190098	190332	+	0
	>2245073 /39872				
25	len =	2135	nex =	4	
	Init	192291	192840	+	0
	Intr	193297	193492	+	0
	Intr	193589	193720	+	0
30	Term	194093	194425	+	0
	>2245073 /6709				
35	len =	1058	nex =	2	
	Term	198909	198442	-	0
	Init	199499	199146	-	0
	>2245073 /94923				
40	len =	739	nex =	2	
	Init	20607	20828	+	0
	Term	20918	21345	+	0
45	>2245073 /24997				
	len =	530	nex =	1	
50	Sngl	26357	25828	-	0
	>2245073 /33509				
55	len =	1450	nex =	2	
	Init	38766	39446	+	0
	Term	39638	40214	+	0
60	>2245073 /35260				

938

	len =	1557	nex =	3	
	Term	43961	43535	-	0
	Intr	44176	44048	-	0
5	Init	45091	44398	-	0
	>2245073 /27500				
10	len =	1700	nex =	1	
	Sngl	51675	51471	-	0
	>2245073 /99796				
15	len =	1150	nex =	2	
	Init	64024	64466	+	0
	Term	64647	65171	+	0
20	>2245073 /31538				
	len =	1278	nex =	4	
	Term	79423	79066	-	0
25	Intr	79725	79528	-	0
	Intr	80213	80047	-	0
	Init	80343	80313	-	0
	>2245073 /26448				
30	len =	2146	nex =	8	
	Term	87600	87509	-	0
	Intr	87818	87699	-	0
35	Intr	88211	88116	-	0
	Intr	88333	88295	-	0
	Intr	88636	88458	-	0
	Intr	88765	88726	-	0
	Intr	88913	88854	-	0
40	Init	89406	89167	-	0
	>2245126 /39922				
45	len =	2134	nex =	5	
	Term	28671	27817	-	0
	Intr	28825	28745	-	0
	Intr	28988	28913	-	0
	Intr	29183	29080	-	0
50	Init	29950	29830	-	0
	>2245126 /37533				
55	len =	1873	nex =	7	
	Init	30483	30887	+	0
	Intr	30977	31070	+	0
	Intr	31153	31292	+	0
	Intr	31365	31439	+	0
60	Intr	31521	31678	+	0

				939	
	Intr	31762	31823	+	0
	Term	31972	32355	+	0
	>2245126	/42815			
5	len =	1514	nex =	4	
	Init	56618	56988	+	0
	Intr	57254	57524	+	0
10	Intr	57621	57791	+	0
	Term	57887	58131	+	0
	>2252639	/36439			
15	len =	2305	nex =	12	
	Term	112752	112679	-	0
	Intr	112953	112837	-	0
	Intr	113158	113042	-	0
20	Intr	113355	113254	-	0
	Intr	113539	113444	-	0
	Intr	113704	113623	-	0
	Intr	113928	113814	-	0
	Intr	114069	114018	-	0
25	Intr	114227	114147	-	0
	Intr	114489	114328	-	0
	Intr	114748	114572	-	0
	Init	114983	114885	-	0
30	>2252639	/32628			
	len =	8176	nex =	7	
	Term	112752	112549	-	0
35	Intr	112953	112837	-	0
	Intr	113158	113042	-	0
	Intr	113355	113254	-	0
	Intr	113539	113444	-	0
	Intr	113704	113623	-	0
40	Init	113928	113814	-	0
	>2252639	/7870			
	len =	2062	nex =	9	
45	Init	55275	55373	+	0
	Intr	55679	55864	+	0
	Intr	55943	56072	+	0
	Intr	56168	56248	+	0
50	Intr	56342	56529	+	0
	Intr	56624	56719	+	0
	Intr	56822	56915	+	0
	Intr	57043	57162	+	0
	Term	57257	57336	+	0
55	>2252639	/42847			
	len =	2459	nex =	3	
60	Init	64066	64204	+	0

					940
	Intr	65296	65804	+	0
	Term	65895	66271	+	0
5	>2252639	/20756			
	len =	561	nex =	1	
	Sngl	66935	66375	-	0
10	>2252639	/8355			
	len =	619	nex =	1	
	Sngl	67016	66406	-	0
15	>2252639	/104398			
	len =	114	nex =	1	
20	Sngl	67655	67768	+	0
	>2252639	/34829			
25	len =	1550	nex =	5	
	Term	72152	71686	-	0
	Intr	72324	72213	-	0
	Intr	72574	72402	-	0
	Intr	72867	72664	-	0
30	Init	73235	73005	-	0
	>2252639	/34276			
35	len =	2157	nex =	6	
	Term	76139	75823	-	0
	Intr	76346	76218	-	0
	Intr	76530	76444	-	0
	Intr	76771	76626	-	0
40	Intr	76952	76898	-	0
	Init	77979	77037	-	0
	>2252639	/11108			
45	len =	539	nex =	1	
	Sngl	79342	78804	-	0
50	>2252639	/1269			
	len =	1433	nex =	3	
	Term	79851	79679	-	0
	Intr	80212	80012	-	0
55	Init	80700	80396	-	0
	>2252639	/5476			
60	len =	835	nex =	3	

				941	
	Init	85064	85271	+	0
	Intr	85376	85455	+	0
	Term	85554	85898	+	0
5	>2252639	/35833			
	len =	873	nex =	3	
10	Init	85064	85271	+	0
	Intr	85376	85455	+	0
	Term	85554	85936	+	0
	>2252639	/1810			
15	len =	878	nex =	3	
	Init	85064	85271	+	0
	Intr	85376	85455	+	0
	Term	85554	85941	+	0
20	>2252639	/17857			
	len =	910	nex =	3	
25	Init	85064	85271	+	0
	Intr	85376	85455	+	0
	Term	85554	85972	+	0
	>2252639	/10862			
30	len =	864	nex =	3	
	Init	85068	85271	+	0
	Intr	85376	85455	+	0
35	Term	85554	85931	+	0
	>2252639	/22773			
	len =	2008	nex =	2	
40	Term	92196	90691	-	0
	Init	92698	92411	-	0
	>2252823	/11106			
45	len =	1289	nex =	1	
	Sngl	107171	108459	+	0
50	>2252823	/25765			
	len =	315	nex =	1	
	Sngl	1671	1357	-	0
55	>2252823	/38970			
	len =	2486	nex =	1	
60	Sngl	29968	30145	+	0

```

>2252823      /15741
5      len =    3070    nex =    3
      Init  29968    30145    +    0
      Intr  30436    30547    +    0
      Term  30642    31104    +    0

10 >2252823      /28637
      len =    2900    nex =    3
      Init  35493    36349    +    0
15      Intr  36852    37326    +    0
      Term  37673    38392    +    0

      >2252823      /21038
20      len =    495    nex =    1
      Sngl  37895    38389    +    0

      >2252823      /35506
25      len =    582    nex =    1
      Sngl  50035    49454    -    0

30 >2252823      /39479
      len =    1604    nex =    3
      Term  56402    56064    -    0
35      Intr  57185    56486    -    0
      Init  57649    57493    -    0

      >2252823      /36326
40      len =    2392    nex =    3
      Term  64455    64054    -    0
      Intr  64734    64625    -    0
45      Init  65205    64824    -    0

      >2252823      /31027
      len =    1150    nex =    2
50      Init  94085    94153    +    0
      Term  94219    95230    +    0

      >2252848      /111719
55      len =    733    nex =    1
      Sngl  46064    45332    -    0

60 >2252848      /11036

```

	len =	790	nex =	1	
	Sngl	46089	45304	-	0
5	>2252848	/3204			
	len =	833	nex =	1	
	Sngl	60597	61429	+	0
10	>2252848	/22161			
	len =	670	nex =	1	
15	Sngl	63070	63731	+	0
	>2252848	/22348			
	len =	740	nex =	1	
20	Sngl	65608	64869	-	0
	>2252848	/28082			
25	len =	1216	nex =	3	
	Init	80915	80991	+	0
	Intr	81337	81552	+	0
	Term	81645	81897	+	0
30	>2252848	/26442			
	len =	1210	nex =	3	
35	Init	80915	80991	+	0
	Intr	81337	81552	+	0
	Term	81645	81895	+	0
	>2252848	/37305			
40	len =	1575	nex =	4	
	Term	91905	91570	-	0
	Intr	92168	92002	-	0
45	Intr	92528	92246	-	0
	Init	92758	92613	-	0
	>2252848	/37175			
50	len =	2050	nex =	3	
	Term	95449	94674	-	0
	Intr	95668	95551	-	0
	Init	96720	96101	-	0
55	>2262097	/22611			
	len =	1439	nex =	4	
60	Init	31	168	+	0

				944	
	Intr	253	403	+	0
	Intr	481	885	+	0
	Term	969	1469	+	0
5	>2262097	/37663			
	len =	1694	nex =	2	
	Term	48814	47723	-	0
10	Init	49413	49234	-	0
	>2262097	/37704			
	len =	1990	nex =	6	
15	Term	4521	4199	-	0
	Intr	4778	4665	-	0
	Intr	5379	5207	-	0
	Intr	5540	5489	-	0
20	Intr	5680	5632	-	0
	Init	6186	5782	-	0
	>2262097	/112955			
25	len =	2350	nex =	4	
	Term	89371	88825	-	0
	Intr	89563	89456	-	0
	Intr	89803	89654	-	0
30	Init	91172	90509	-	0
	>2262135	/41490			
	len =	1454	nex =	2	
35	Term	2318	1916	-	0
	Init	3369	2625	-	0
	>2262135	/20167			
40	len =	1304	nex =	2	
	Term	4241	3765	-	0
	Init	5068	4768	-	0
45	>2262135	/32291			
	len =	1390	nex =	3	
50	Term	3887	3685	-	0
	Intr	4241	4100	-	0
	Init	5072	4768	-	0
	>2262135	/6568			
55	len =	2212	nex =	6	
	Term	55501	55152	-	0
	Intr	55716	55591	-	0
60	Intr	55868	55793	-	0

				945
	Intr	56088	55950	- 0
	Intr	56564	56483	- 0
	Init	57080	56653	- 0
5	>2262135	/10207		
	len =	2063	nex =	4
	Init	59951	60024	+ 0
10	Intr	60681	60762	+ 0
	Intr	61016	61098	+ 0
	Term	61517	61813	+ 0
	>2262135	/18545		
15	len =	647	nex =	1
	Sngl	6145	6791	+ 0
20	>2262135	/4346		
	len =	2939	nex =	6
	Init	70603	71150	+ 0
25	Intr	71555	71677	+ 0
	Intr	71842	71907	+ 0
	Intr	71994	72059	+ 0
	Intr	72734	72814	+ 0
	Term	72893	73541	+ 0
30	>2262135	/26127		
	len =	817	nex =	1
35	Sngl	10051	10199	+ 0
	>2262135	/8114		
	len =	1879	nex =	4
40	Init	97068	97416	+ 0
	Intr	98158	98297	+ 0
	Intr	98468	98540	+ 0
	Term	98650	98946	+ 0
45	>2262135	/34186		
	len =	347	nex =	1
50	Sngl	97069	97415	+ 0
	>2262135	/145375		
	len =	319	nex =	1
55	Sngl	10051	10164	+ 0
	>2262135	/18454		
60	len =	354	nex =	1

	Sngl	10051	10199	+	0
	>2262135	/27915			
5	len =	1173	nex =	3	
	Init	99470	99712	+	0
	Intr	99822	99870	+	0
10	Term	99982	100642	+	0
	>2262155	/1441			
	len =	657	nex =	1	
15	Sngl	23119	22463	-	0
	>2262155	/38365			
20	len =	2443	nex =	11	
	Term	33741	33609	-	0
	Intr	33874	33812	-	0
	Intr	34038	33961	-	0
25	Intr	34207	34130	-	0
	Intr	34357	34283	-	0
	Intr	34542	34456	-	0
	Intr	35004	34864	-	0
	Intr	35174	35106	-	0
30	Intr	35320	35254	-	0
	Intr	35536	35471	-	0
	Init	36051	35849	-	0
	>2262155	/2578			
35	len =	2710	nex =	12	
	Term	41819	41536	-	0
	Intr	42007	41945	-	0
40	Intr	42177	42100	-	0
	Intr	42353	42276	-	0
	Intr	42507	42433	-	0
	Intr	42691	42605	-	0
	Intr	42920	42792	-	0
45	Intr	43144	43004	-	0
	Intr	43300	43232	-	0
	Intr	43448	43382	-	0
	Intr	43690	43625	-	0
	Init	44238	44044	-	0
50	>2262155	/10042			
	len =	1776	nex =	4	
55	Init	47118	47195	+	0
	Intr	47279	47459	+	0
	Intr	47575	47672	+	0
	Term	47837	48384	+	0
60	>2262155	/13246			

	len =	1990	nex =	6	
	Init	54079	54165	+	0
5	Intr	54255	54346	+	0
	Intr	54432	54540	+	0
	Intr	54640	54675	+	0
	Intr	54764	54850	+	0
	Term	54940	55113	+	0
10	>2262155 /34698				
	len =	1459	nex =	6	
15	Init	56211	56260	+	0
	Intr	56344	56556	+	0
	Intr	56654	56802	+	0
	Intr	56878	57034	+	0
	Intr	57160	57252	+	0
20	Term	57530	57669	+	0
	>2262155 /39211				
	len =	2110	nex =	2	
25	Init	64477	65546	+	0
	Term	66273	66579	+	0
	>2262155 /19601				
30	len =	2050	nex =	2	
	Init	64534	65546	+	0
	Term	66273	66579	+	0
35	>2262155 /32751				
	len =	850	nex =	1	
40	Sngl	77445	76604	-	0
	>2262155 /3276				
	len =	1167	nex =	1	
45	Sngl	8628	9794	+	0
	>2264302 /38370				
50	len =	1450	nex =	2	
	Term	35101	34004	-	0
	Init	35452	35188	-	0
55	>2264302 /9562				
	len =	2074	nex =	0	
60	>2264302 /28046				

948

	len =	1581	nex =	4	
	Term	51719	51257	-	0
	Intr	52040	51910	-	0
5	Intr	52474	52402	-	0
	Init	52837	52724	-	0
	>2264302 /16428				
10	len =	1571	nex =	4	
	Term	51818	51294	-	0
	Intr	52040	51910	-	0
	Intr	52474	52402	-	0
15	Init	52864	52724	-	0
	>2264302 /100085				
20	len =	1254	nex =	3	
	Term	5287	4881	-	0
	Intr	5613	5357	-	0
	Init	6134	5782	-	0
25	>2264303 /22				
	len =	1735	nex =	6	
	Init	14289	14642	+	0
30	Intr	14799	14910	+	0
	Intr	15002	15095	+	0
	Intr	15228	15405	+	0
	Intr	15488	15557	+	0
35	Term	15638	16023	+	0
	>2264303 /7145				
	len =	824	nex =	4	
40	Init	3387	3465	+	0
	Intr	3544	3666	+	0
	Intr	3754	3870	+	0
	Term	3947	4205	+	0
45	>2264303 /4273				
	len =	1845	nex =	3	
	Term	45044	44650	-	0
50	Intr	45266	45126	-	0
	Init	46494	46178	-	0
	>2264303 /35612				
55	len =	1469	nex =	4	
	Init	58748	59002	+	0
	Intr	59229	59277	+	0
	Intr	59634	59833	+	0
60	Term	59930	60216	+	0

```

>2264303      /42336

    len =      1825    nex =      4
5
    Term    64023    63682      -      0
    Intr    64570    64473      -      0
    Intr    65089    64989      -      0
    Init    65506    65289      -      0
10
>2264304      /34402

    len =      2558    nex =      5
15
    Init    20281    20902      +      0
    Intr    21285    21510      +      0
    Intr    21627    21849      +      0
    Intr    22104    22317      +      0
    Term    22554    22838      +      0
20
>2264304      /34783

    len =      2075    nex =      5
25
    Term    23983    23714      -      0
    Intr    24174    24080      -      0
    Intr    24709    24267      -      0
    Intr    25149    24793      -      0
    Init    25788    25400      -      0
30
>2264304      /39319

    len =      1870    nex =      5
35
    Init    2871     2989      +      0
    Intr    3690     3771      +      0
    Intr    3960     4165      +      0
    Intr    4328     4381      +      0
    Term    4476     4733      +      0
40
>2264304      /9159

    len =      1570    nex =      2
45
    Init    41803    42064      +      0
    Term    42974    43372      +      0

>2264304      /38464

50
    len =      1270    nex =      1

    Sngl    51034    52303      +      0

>2264304      /28578

55
    len =      2110    nex =      5

    Init     515     1139      +      0
    Intr    1407     1504      +      0
60
    Intr    1754     1853      +      0

```

				950	
	Intr	2027	2272	+	0
	Term	2358	2618	+	0
5	>2264304	/41195			
	len =	353	nex =	1	
	Sngl	57898	57549	-	0
10	>2264304	/2871			
	len =	430	nex =	2	
15	Init	6595	6647	+	0
	Term	6733	7019	+	0
	>2264304	/30073			
20	len =	1810	nex =	1	
	Sngl	65320	65030	-	0
	>2264304	/32071			
25	len =	1128	nex =	1	
	Sngl	67814	67283	-	0
30	>2264304	/103464			
	len =	1096	nex =	1	
	Sngl	67814	67316	-	0
35	>2264304	/17818			
	len =	1136	nex =	1	
40	Sngl	67814	67277	-	0
	>2264304	/24095			
	len =	596	nex =	1	
45	Sngl	72223	72818	+	0
	>2264304	/111741			
50	len =	2898	nex =	9	
	Init	77610	77692	+	0
	Intr	78044	78153	+	0
	Intr	78600	78734	+	0
	Intr	78876	79022	+	0
55	Intr	79400	79483	+	0
	Intr	79589	79635	+	0
	Intr	79729	79802	+	0
	Intr	79915	79973	+	0
60	Term	80152	80212	+	0

	>2264305	/10263			
	len =	1493	nex =	5	
5	Init	31119	31386	+	0
	Intr	31604	31784	+	0
	Intr	31864	32005	+	0
	Intr	32090	32159	+	0
	Term	32249	32611	+	0
10	>2264305	/98400			
	len =	993	nex =	3	
15	Term	4415	4173	-	0
	Intr	4868	4742	-	0
	Init	5152	4965	-	0
20	>2264305	/36333			
	len =	1450	nex =	4	
	Term	4415	4119	-	0
	Intr	4868	4742	-	0
25	Intr	5244	4965	-	0
	Init	5422	5374	-	0
	>2264305	/121728			
30	len =	550	nex =	2	
	Term	5244	5080	-	0
	Init	5422	5374	-	0
35	>2264305	/41072			
	len =	1312	nex =	4	
	Term	4415	4326	-	0
40	Intr	4868	4742	-	0
	Intr	5244	4965	-	0
	Init	5422	5374	-	0
45	>2264305	/24983			
	len =	599	nex =	1	
	Sngl	64677	64079	-	0
50	>2264305	/16865			
	len =	1615	nex =	4	
	Init	71009	71096	+	0
55	Intr	71447	71574	+	0
	Intr	71737	71841	+	0
	Term	72035	72347	+	0
60	>2264305	/35698			

952

	len =	1150	nex =	4	
	Init	71025	71096	+	0
	Intr	71447	71574	+	0
5	Intr	71737	71841	+	0
	Term	72035	72162	+	0
	>2264306 /21505				
10	len =	1450	nex =	3	
	Term	10517	10132	-	0
	Intr	11048	10721	-	0
	Init	11577	11269	-	0
15	>2264306 /19024				
	len =	715	nex =	2	
20	Term	14439	14066	-	0
	Init	14777	14527	-	0
	>2264306 /33140				
25	len =	1450	nex =	3	
	Term	14439	13966	-	0
	Intr	14854	14527	-	0
	Init	15411	14979	-	0
30	>2264306 /121213				
	len =	333	nex =	1	
35	Sngl	2596	2928	+	0
	>2264306 /39888				
40	len =	2203	nex =	9	
	Term	35099	34644	-	0
	Intr	35279	35181	-	0
	Intr	35475	35371	-	0
	Intr	35651	35559	-	0
45	Intr	35855	35763	-	0
	Intr	36011	35958	-	0
	Intr	36218	36117	-	0
	Intr	36369	36295	-	0
	Init	36846	36503	-	0
50	>2264306 /11054				
	len =	1417	nex =	5	
55	Init	41110	41228	+	0
	Intr	41333	41424	+	0
	Intr	41763	41818	+	0
	Intr	42120	42181	+	0
	Term	42324	42526	+	0
60					

	>2264306	/3699			
	len =	1897	nex =	4	
5	Init	5030	5266	+	0
	Intr	5420	6238	+	0
	Intr	6325	6526	+	0
	Term	6551	6926	+	0
10	>2264306	/6637			
	len =	1428	nex =	3	
	Term	80382	79690	-	0
15	Intr	80764	80484	-	0
	Init	81117	80852	-	0
	>2264306	/111669			
20	len =	382	nex =	2	
	Init	88535	88581	+	0
	Term	88664	88916	+	0
25	>2264307	/42441			
	len =	682	nex =	2	
	Term	48650	48344	-	0
30	Init	49017	48966	-	0
	>2264307	/22848			
	len =	658	nex =	2	
35	Term	48650	48368	-	0
	Init	49017	48966	-	0
	>2264307	/145394			
40	len =	638	nex =	2	
	Term	48650	48388	-	0
	Init	49017	48966	-	0
45	>2264307	/11511			
	len =	776	nex =	2	
50	Term	48650	48252	-	0
	Init	49027	48966	-	0
	>2264307	/12330			
55	len =	670	nex =	2	
	Term	48650	48363	-	0
	Init	49017	48966	-	0
60	>2264307	/37668			

	len =	2959	nex =	12	
	Term	58676	58435	-	0
5	Intr	58819	58762	-	0
	Intr	59006	58939	-	0
	Intr	59148	59089	-	0
	Intr	59415	59374	-	0
	Intr	59547	59504	-	0
10	Intr	59753	59684	-	0
	Intr	60223	60104	-	0
	Intr	60499	60481	-	0
	Intr	60688	60616	-	0
	Intr	60911	60847	-	0
15	Init	61393	61056	-	0
>2264307		/24058			
	len =	1653	nex =	4	
20	Init	72492	72816	+	0
	Intr	73287	73411	+	0
	Intr	73485	73593	+	0
	Term	73888	74144	+	0
25	>2264308		/1935		
	len =	1396	nex =	1	
30	Sngl	17599	16204	-	0
>2264308		/22483			
	len =	2981	nex =	8	
35	Term	4792	4416	-	0
	Intr	5296	4866	-	0
	Intr	5495	5375	-	0
	Intr	5737	5588	-	0
40	Intr	6028	5823	-	0
	Intr	6224	6110	-	0
	Intr	6544	6307	-	0
	Init	7396	7131	-	0
45	>2264309		/37959		
	len =	357	nex =	1	
	Sngl	16800	16444	-	0
50	>2264309		/15155		
	len =	872	nex =	2	
55	Init	22581	22830	+	0
	Term	22927	23337	+	0
>2264309		/36334			
60	len =	4030	nex =	8	

955

	Term	23729	23461	-	0
	Intr	23957	23827	-	0
	Intr	24155	24049	-	0
5	Intr	24319	24241	-	0
	Intr	24499	24413	-	0
	Intr	26484	26236	-	0
	Intr	26721	26572	-	0
	Init	27488	26913	-	0
10	>2264309 /109246				
	len =	614	nex =	1	
15	Sngl	36598	37211	+	0
	>2264309 /34868				
	len =	2755	nex =	10	
20	Init	56456	56771	+	0
	Intr	57170	57262	+	0
	Intr	57346	57427	+	0
	Intr	57612	57708	+	0
25	Intr	57802	57877	+	0
	Intr	58009	58067	+	0
	Intr	58236	58358	+	0
	Intr	58523	58580	+	0
	Intr	58667	58752	+	0
30	Term	58834	59210	+	0
	>2264310 /99461				
	len =	692	nex =	1	
35	Sngl	11215	11906	+	0
	>2264310 /15761				
40	len =	2548	nex =	6	
	Term	19001	18686	-	0
	Intr	19291	19099	-	0
	Intr	19675	19440	-	0
45	Intr	19965	19793	-	0
	Intr	20557	20507	-	0
	Init	21233	20635	-	0
	>2264310 /11083				
50	len =	565	nex =	1	
	Sngl	2390	2039	-	0
55	>2264310 /31527				
	len =	589	nex =	1	
60	Sngl	45291	45879	+	0

	>2264310	/17408		
	len =	642	nex =	1
5	Sngl	75188	74547	- 0
	>2264310	/125083		
	len =	1961	nex =	5
10	Init	8184	8440	+ 0
	Intr	8574	8786	+ 0
	Intr	8879	9037	+ 0
	Intr	9616	9684	+ 0
15	Term	9797	10144	+ 0
	>2264311	/32868		
	len =	1724	nex =	5
20	Term	22845	22268	- 0
	Intr	23036	22924	- 0
	Intr	23230	23115	- 0
	Intr	23684	23307	- 0
25	Init	23977	23868	- 0
	>2264311	/6256		
	len =	970	nex =	3
30	Term	61688	61655	- 0
	Intr	61915	61777	- 0
	Init	62223	62000	- 0
35	>2264311	/125951		
	len =	2213	nex =	8
	Term	60708	60456	- 0
40	Intr	60920	60814	- 0
	Intr	61074	61009	- 0
	Intr	61491	61410	- 0
	Intr	61688	61644	- 0
	Intr	61915	61777	- 0
45	Intr	62223	62000	- 0
	Init	62668	62430	- 0
	>2264311	/27195		
50	len =	1880	nex =	7
	Term	82920	82401	- 0
	Intr	83150	83009	- 0
	Intr	83482	83243	- 0
55	Intr	83616	83581	- 0
	Intr	83788	83708	- 0
	Intr	83928	83871	- 0
	Init	84280	84011	- 0
60	>2264312	/14950		

	len =	881	nex =	1	
5	Sngl	27808	26928	-	0
	>2264312	/95433			
	len =	1318	nex =	5	
10	Term	41828	41661	-	0
	Intr	42031	41958	-	0
	Intr	42285	42119	-	0
	Intr	42519	42450	-	0
	Init	42741	42601	-	0
15	>2264312	/41937			
	len =	412	nex =	1	
20	Sngl	46315	45915	-	0
	>2264312	/13715			
	len =	505	nex =	1	
25	Sngl	46419	45915	-	0
	>2264312	/20908			
30	len =	1588	nex =	1	
	Sngl	47047	45915	-	0
	>2264312	/121153			
35	len =	1599	nex =	0	
	>2264312	/21872			
40	len =	1999	nex =	5	
	Init	76178	76439	+	0
	Intr	76875	77278	+	0
	Intr	77349	77609	+	0
45	Intr	77680	77802	+	0
	Term	77884	78176	+	0
	>2264312	/40252			
50	len =	929	nex =	3	
	Init	8129	8281	+	0
	Intr	8374	8529	+	0
	Term	8834	9057	+	0
55	>2264313	/13012			
	len =	2530	nex =	3	
60	Init	50735	51416	+	0

					958
	Intr	51723	52053	+	0
	Term	52969	53262	+	0
	>2264313	/156373			
5	len =	1597	nex =	4	
	Term	56197	55946	-	0
	Intr	56442	56319	-	0
10	Intr	57210	56988	-	0
	Init	57542	57464	-	0
	>2264314	/8635			
15	len =	1886	nex =	4	
	Term	10067	9103	-	0
	Intr	10250	10148	-	0
	Intr	10433	10340	-	0
20	Init	10988	10835	-	0
	>2264314	/115644			
25	len =	1259	nex =	2	
	Term	26540	26126	-	0
	Init	27384	26837	-	0
	>2264314	/38996			
30	len =	2313	nex =	7	
	Term	27833	27526	-	0
	Intr	28049	27984	-	0
35	Intr	28349	28278	-	0
	Intr	28813	28492	-	0
	Intr	29046	28886	-	0
	Intr	29175	29131	-	0
	Init	29838	29580	-	0
40	>2264314	/32785			
	len =	1499	nex =	1	
45	Sngl	41738	42167	+	0
	>2264314	/20245			
	len =	1429	nex =	1	
50	Sngl	41738	42147	+	0
	>2264314	/5592			
55	len =	1450	nex =	0	
	>2264314	/13819			
60	len =	1390	nex =	1	

					959
	Sngl	41738	42167	+	0
	>2264314	/29726			
5	len =	673	nex =	1	
	Sngl	46055	46727	+	0
	>2264314	/41900			
10	len =	567	nex =	1	
	Sngl	46131	46697	+	0
15	>2264314	/2462			
	len =	570	nex =	1	
	Sngl	46131	46700	+	0
20	>2264314	/16750			
	len =	585	nex =	1	
25	Sngl	46131	46715	+	0
	>2264314	/18232			
	len =	1571	nex =	5	
30	Term	48315	47879	-	0
	Intr	48456	48413	-	0
	Intr	48598	48541	-	0
	Intr	48919	48826	-	0
35	Init	49449	49182	-	0
	>2264314	/9012			
	len =	1870	nex =	0	
40	>2264314	/7365			
	len =	1776	nex =	0	
45	>2264314	/33059			
	len =	2811	nex =	7	
	Term	61633	61320	-	0
50	Intr	61973	61823	-	0
	Intr	62227	62054	-	0
	Intr	62409	62320	-	0
	Intr	62646	62576	-	0
	Intr	63811	62772	-	0
55	Init	64130	63836	-	0
	>2264314	/27647			
	len =	1370	nex =	3	
60					

				960	
	Init	72212	72591	+	0
	Intr	72849	73086	+	0
	Term	73196	73581	+	0
5	>2264315	/10218			
	len =	2270	nex =	4	
	Term	26015	25438	-	0
10	Intr	26141	26094	-	0
	Intr	27175	26240	-	0
	Init	27707	27384	-	0
	>2264315	/29462			
15	len =	1139	nex =	2	
	Init	45117	45873	+	0
	Term	45961	46255	+	0
20	>2264315	/14965			
	len =	430	nex =	1	
	Sngl	47036	46610	-	0
25	>2264315	/114307			
	len =	464	nex =	1	
30	Sngl	47105	46642	-	0
	>2264315	/3363			
	len =	636	nex =	1	
35	Sngl	47111	46476	-	0
	>2264315	/41666			
40	len =	2157	nex =	12	
	Init	59476	59703	+	0
	Intr	59800	59887	+	0
	Intr	60015	60074	+	0
45	Intr	60160	60192	+	0
	Intr	60278	60355	+	0
	Intr	60433	60476	+	0
	Intr	60582	60622	+	0
	Intr	60709	60791	+	0
50	Intr	60876	60967	+	0
	Intr	61055	61124	+	0
	Intr	61205	61246	+	0
	Term	61348	61632	+	0
55	>2264316	/31759			
	len =	1810	nex =	3	
	Term	40887	40024	-	0
60	Intr	41245	40976	-	0

					961
	Init	41826	41375	-	0
	>2264316	/4716			
5	len =	1150	nex =	4	
	Term	48078	47771	-	0
	Intr	48347	48169	-	0
	Intr	48549	48448	-	0
10	Init	48918	48760	-	0
	>2264316	/35357			
15	len =	3430	nex =	2	
	Init	4937	5508	+	0
	Term	7116	8360	+	0
	>2264316	/13418			
20	len =	1121	nex =	4	
	Term	50134	49841	-	0
	Intr	50452	50271	-	0
25	Intr	50665	50567	-	0
	Init	50961	50832	-	0
	>2264316	/25839			
30	len =	1733	nex =	4	
	Term	52037	51717	-	0
	Intr	52799	52621	-	0
	Intr	52994	52893	-	0
35	Init	53449	53248	-	0
	>2264316	/5103			
40	len =	566	nex =	2	
	Term	56108	55749	-	0
	Init	56314	56188	-	0
	>2264316	/25723			
45	len =	1118	nex =	3	
	Init	70502	70609	+	0
	Intr	70687	70765	+	0
50	Term	71265	71619	+	0
	>2264316	/28686			
55	len =	1761	nex =	5	
	Init	73159	73478	+	0
	Intr	73823	73864	+	0
	Intr	74151	74238	+	0
	Intr	74355	74436	+	0
60	Term	74532	74919	+	0

>2264316 /33187

5	len =	1316	nex =	6	
	Init	75294	75411	+	0
	Intr	75493	75533	+	0
	Intr	75623	75723	+	0
	Intr	75977	76121	+	0
10	Intr	76215	76304	+	0
	Term	76389	76609	+	0

>2264316 /40559

15	len =	940	nex =	4	
	Init	75623	75723	+	0
	Intr	75977	76121	+	0
	Intr	76215	76304	+	0
20	Term	76389	76430	+	0

>2264317 /27304

25	len =	1450	nex =	4	
	Init	10536	10865	+	0
	Intr	11094	11307	+	0
	Intr	11430	11575	+	0
	Term	11678	11977	+	0

>2264317 /41386

30	len =	2230	nex =	7	
35	Init	18624	18806	+	0
	Intr	19320	19433	+	0
	Intr	19544	19688	+	0
	Intr	19786	19863	+	0
	Intr	19964	20076	+	0
40	Intr	20166	20269	+	0
	Term	20364	20848	+	0

>2264317 /19638

45	len =	1116	nex =	5	
	Term	39626	39380	-	0
	Intr	39837	39741	-	0
	Intr	39994	39932	-	0
50	Intr	40263	40110	-	0
	Init	40495	40353	-	0

>2264317 /6734

55	len =	2230	nex =	8	
	Init	43041	43121	+	0
	Intr	43615	43732	+	0
	Intr	43820	43927	+	0
60	Intr	44029	44153	+	0

963

	Intr	44256	44520	+	0
	Intr	44612	44680	+	0
	Intr	44773	44934	+	0
	Term	45031	45269	+	0
5	>2264318 /3797				
	len =	3010	nex =	11	
10	Term	14549	14209	-	0
	Intr	14698	14642	-	0
	Intr	14911	14777	-	0
	Intr	15084	15004	-	0
	Intr	15230	15162	-	0
15	Intr	15408	15334	-	0
	Intr	15837	15757	-	0
	Intr	16050	15932	-	0
	Intr	16304	16139	-	0
	Intr	16522	16393	-	0
20	Init	17210	16609	-	0
	>2264318 /33231				
	len =	1510	nex =	4	
25	Term	19006	18683	-	0
	Intr	19387	19102	-	0
	Intr	19635	19485	-	0
	Init	20191	19835	-	0
30	>2264318 /42276				
	len =	681	nex =	1	
35	Sngl	24794	24114	-	0
	>2264318 /26752				
	len =	754	nex =	1	
40	Sngl	6372	6627	+	0
	>2264318 /25855				
45	len =	2410	nex =	4	
	Init	74093	74435	+	0
	Intr	74770	74907	+	0
	Intr	75288	75359	+	0
50	Term	75730	76502	+	0
	>2264319 /37985				
	len =	1041	nex =	3	
55	Term	29497	28961	-	0
	Intr	29867	29820	-	0
	Init	30001	29945	-	0
60	>2264320 /36697				

	len =	3070	nex =	12	
	Init	77774	77917	+	0
5	Intr	78577	78735	+	0
	Intr	78827	78886	+	0
	Intr	79001	79047	+	0
	Intr	79159	79212	+	0
	Intr	79302	79479	+	0
10	Intr	79602	79754	+	0
	Intr	79848	79913	+	0
	Intr	80000	80047	+	0
	Intr	80127	80233	+	0
	Intr	80327	80405	+	0
15	Term	80513	80843	+	0
>2264321 /22350					
	len =	1378	nex =	1	
20	Sngl	42485	43862	+	0
>2264321 /17195					
25	len =	2639	nex =	9	
	Term	44138	43885	-	0
	Intr	44545	44459	-	0
	Intr	44723	44638	-	0
30	Intr	45000	44910	-	0
	Intr	45164	45079	-	0
	Intr	45369	45261	-	0
	Intr	45565	45519	-	0
	Intr	45725	45656	-	0
35	Init	46523	45984	-	0
>2264321 /4025					
	len =	1845	nex =	8	
40	Term	62897	62608	-	0
	Intr	63027	62981	-	0
	Intr	63329	63108	-	0
	Intr	63461	63409	-	0
45	Intr	63720	63555	-	0
	Intr	63853	63812	-	0
	Intr	64024	63940	-	0
	Init	64452	64119	-	0
50	>2264321 /226				
	len =	1214	nex =	3	
	Init	64562	64824	+	0
55	Intr	65227	65327	+	0
	Term	65506	65775	+	0
>2264321 /25843					
60	len =	1179	nex =	1	

	Sngl	64602	65780	+	0
	>2264367	/13226			
5	len =	760	nex =	1	
	Sngl	17702	16945	-	0
10	>2264367	/6280			
	len =	1721	nex =	4	
	Init	79635	80401	+	0
15	Intr	80649	80739	+	0
	Intr	80875	81047	+	0
	Term	81136	81355	+	0
	>2264367	/14253			
20	len =	394	nex =	1	
	Sngl	79694	80087	+	0
25	>2264367	/2093			
	len =	1697	nex =	4	
	Term	81924	81450	-	0
30	Intr	82092	82014	-	0
	Intr	82411	82172	-	0
	Init	83146	82545	-	0
	>2275194	/35109			
35	len =	1541	nex =	0	
	>2275194	/20378			
40	len =	540	nex =	1	
	Sngl	46427	45888	-	0
	>2275194	/6324			
45	len =	564	nex =	1	
	Sngl	81129	81692	+	0
50	>2275194	/95662			
	len =	550	nex =	0	
	>2275194	/21715			
55	len =	339	nex =	1	
	Sngl	81340	81678	+	0
60	>2275194	/34414			

	len =	2177	nex =	6	
	Init	1529	1687	+	0
5	Intr	1807	1877	+	0
	Intr	2195	2314	+	0
	Intr	2406	2524	+	0
	Intr	2616	2697	+	0
	Term	2789	3076	+	0
10	>2275194 /35584				
	len =	2064	nex =	6	
15	Init	1529	1687	+	0
	Intr	1807	1877	+	0
	Intr	2195	2314	+	0
	Intr	2406	2524	+	0
	Intr	2616	2697	+	0
20	Term	2789	3017	+	0
	>2281081 /99937				
	len =	1179	nex =	4	
25	Init	17994	18277	+	0
	Intr	18570	18617	+	0
	Intr	18757	18836	+	0
	Term	18973	19172	+	0
30	>2281081 /34407				
	len =	1614	nex =	2	
35	Term	20892	20042	-	0
	Init	21655	20980	-	0
	>2281081 /24415				
40	len =	1398	nex =	1	
	Sngl	37217	37695	+	0
	>2281081 /21866				
45	len =	3043	nex =	10	
	Init	41405	41802	+	0
	Intr	41989	42098	+	0
50	Intr	42186	42243	+	0
	Intr	42347	42610	+	0
	Intr	42881	43018	+	0
	Intr	43151	43202	+	0
	Intr	43288	43367	+	0
55	Intr	43475	43534	+	0
	Intr	43663	43743	+	0
	Term	44186	44447	+	0
60	>2281081 /117763				

967

	len =	168	nex =	1	
	Sngl	44280	44447	+	0
5	>2281081	/37969			
	len =	1630	nex =	2	
10	Init	45636	46252	+	0
	Term	46437	47256	+	0
	>2281081	/97249			
15	len =	1570	nex =	3	
	Init	75474	75567	+	0
	Intr	75664	75773	+	0
	Term	76110	76381	+	0
20	>2288979	/30737			
	len =	1957	nex =	6	
25	Term	23749	23549	-	0
	Intr	24382	24215	-	0
	Intr	24583	24465	-	0
	Intr	24734	24673	-	0
	Intr	24906	24830	-	0
30	Init	25505	25278	-	0
	>2288979	/42038			
	len =	2417	nex =	6	
35	Term	26123	25700	-	0
	Intr	26352	26213	-	0
	Intr	26728	26523	-	0
	Intr	27113	27007	-	0
	Intr	27509	27330	-	0
40	Init	28116	27832	-	0
	>2288979	/5460			
45	len =	1369	nex =	2	
	Term	61213	60939	-	0
	Init	61831	61648	-	0
50	>2288979	/31535			
	len =	971	nex =	1	
	Sngl	6953	5983	-	0
55	>2288979	/15927			
	len =	582	nex =	2	
60	Init	83467	83577	+	0
	Term	83732	84048	+	0

```

>2288979      /14769

5      len =      598      nex =      2
      Init  84968      85307      +      0
      Term  85318      85559      +      0

10     >2288979      /22360
      len =      621      nex =      3
      Init  84968      85076      +      0
      Intr  85239      85307      +      0
15     Term  85318      85582      +      0

      >2288979      /8155
20     len =      637      nex =      3
      Init  84968      85076      +      0
      Intr  85239      85307      +      0
      Term  85318      85598      +      0

25     >2288979      /91704
      len =      685      nex =      2
      Init  85879      86099      +      0
30     Term  86237      86563      +      0

      >2288979      /14241
35     len =      594      nex =      2
      Init  85971      86099      +      0
      Term  86237      86564      +      0

40     >2288979      /27364
      len =      593      nex =      1
      Sngl  87277      87869      +      0

45     >2288979      /16079
      len =      558      nex =      2
      Init  88140      88265      +      0
50     Term  88343      88697      +      0

      >2288979      /85
55     len =      670      nex =      2
      Init  88140      88265      +      0
      Term  88343      88809      +      0

60     >2288979      /35036

```

	len =	510	nex =	2	
	Term	89987	89866	-	0
	Init	90375	90255	-	0
5	>2326340 /17730				
	len =	938	nex =	3	
10	Init	12848	12929	+	0
	Intr	13222	13294	+	0
	Term	13456	13785	+	0
	>2335089 /17415				
15	len =	1711	nex =	2	
	Init	18997	19833	+	0
	Term	20359	20707	+	0
20	>2335089 /41462				
	len =	2561	nex =	7	
25	Init	77553	77859	+	0
	Intr	78200	78282	+	0
	Intr	78527	78615	+	0
	Intr	78796	78869	+	0
	Intr	78950	79000	+	0
30	Intr	79347	79408	+	0
	Term	79492	80113	+	0
	>2337888 /30632				
35	len =	597	nex =	1	
	Sngl	45399	44803	-	0
	>2337888 /33132				
40	len =	1427	nex =	1	
	Sngl	56372	54946	-	0
45	>2337888 /25271				
	len =	1190	nex =	4	
	Term	81979	81460	-	0
50	Intr	82251	82069	-	0
	Intr	82443	82348	-	0
	Init	82649	82529	-	0
	>2337888 /36364				
55	len =	2473	nex =	10	
	Term	81979	81474	-	0
	Intr	82251	82069	-	0
60	Intr	82443	82348	-	0

				970	
	Intr	82588	82529	-	0
	Intr	82726	82673	-	0
	Intr	82906	82829	-	0
	Intr	83042	82989	-	0
5	Intr	83230	83147	-	0
	Intr	83655	83627	-	0
	Init	83946	83768	-	0
	>2337888	/48			
10	len =	139	nex =	1	
	Sngl	84105	83967	-	0
15	>2337888	/39291			
	len =	1330	nex =	2	
	Init	9724	10277	+	0
20	Term	10380	11048	+	0
	>2341023	/20848			
	len =	2394	nex =	8	
25	Term	105381	105134	-	0
	Intr	105770	105531	-	0
	Intr	106011	105948	-	0
	Intr	106356	106242	-	0
30	Intr	106669	106531	-	0
	Intr	106971	106841	-	0
	Intr	107209	107080	-	0
	Init	107527	107476	-	0
35	>2341023	/4513			
	len =	1287	nex =	3	
	Term	16071	15822	-	0
40	Intr	16960	16676	-	0
	Init	17108	17082	-	0
	>2341023	/26558			
45	len =	1150	nex =	3	
	Term	23857	23331	-	0
	Intr	24045	23945	-	0
	Init	24472	24392	-	0
50	>2341023	/23398			
	len =	2892	nex =	2	
55	Term	36567	36137	-	0
	Init	39028	38927	-	0
	>2341023	/40467			
60	len =	2202	nex =	7	

971

	Init	41815	41979	+	0
	Intr	42299	42457	+	0
	Intr	42564	42739	+	0
5	Intr	42897	43174	+	0
	Intr	43264	43399	+	0
	Intr	43492	43603	+	0
	Term	43692	44016	+	0
10	>2341023	/19832			
	len =	2656	nex =	4	
	Init	45198	45615	+	0
15	Intr	45720	45944	+	0
	Intr	46040	46752	+	0
	Term	46898	47342	+	0
	>2341023	/91880			
20	len =	1118	nex =	2	
	Init	46306	46752	+	0
	Term	46898	47423	+	0
25	>2341023	/8374			
	len =	805	nex =	3	
30	Init	84788	85031	+	0
	Intr	85113	85256	+	0
	Term	85340	85592	+	0
	>2341023	/9471			
35	len =	649	nex =	1	
	Sngl	85423	85236	-	0
40	>2341023	/30909			
	len =	1020	nex =	3	
	Init	90351	90483	+	0
45	Intr	90571	90628	+	0
	Term	91104	91353	+	0
	>2341023	/28606			
50	len =	730	nex =	1	
	Sngl	91839	92568	+	0
	>2341023	/125151			
55	len =	310	nex =	1	
	Sngl	96904	96600	-	0
60	>2341023	/33613			

	len =	2290	nex =	5	
	Term	94901	94658	-	0
5	Intr	95464	95403	-	0
	Intr	95744	95606	-	0
	Intr	96270	96059	-	0
	Init	96946	96584	-	0
10	>2342673	/21644			
	len =	568	nex =	2	
	Init	1	19	+	0
15	Term	287	568	+	0
	>2342673	/4236			
	len =	1031	nex =	1	
20	Sngl	15499	14469	-	0
	>2342673	/13218			
25	len =	600	nex =	1	
	Sngl	59777	59178	-	0
	>2342673	/1911			
30	len =	1410	nex =	0	
	>2342673	/15745			
35	len =	2693	nex =	15	
	Term	72951	72598	-	0
	Intr	73173	73059	-	0
	Intr	73327	73268	-	0
40	Intr	73473	73420	-	0
	Intr	73651	73592	-	0
	Intr	73809	73747	-	0
	Intr	73936	73893	-	0
	Intr	74109	74025	-	0
45	Intr	74283	74203	-	0
	Intr	74471	74379	-	0
	Intr	74618	74554	-	0
	Intr	74789	74714	-	0
	Intr	74956	74891	-	0
50	Intr	75176	75051	-	0
	Init	75290	75255	-	0
	>2342673	/20814			
55	len =	2669	nex =	14	
	Term	87698	87414	-	0
	Intr	87906	87792	-	0
	Intr	88057	87998	-	0
60	Intr	88219	88166	-	0

973

	Intr	88375	88316	-	0
	Intr	88529	88467	-	0
	Intr	88664	88621	-	0
	Intr	88853	88769	-	0
5	Intr	89044	88964	-	0
	Intr	89241	89149	-	0
	Intr	89408	89344	-	0
	Intr	89583	89508	-	0
	Intr	89751	89686	-	0
10	Init	89916	89851	-	0

>2342673 /36585

15	len =	3206	nex =	16	
	Term	87698	87522	-	0
	Intr	87906	87792	-	0
	Intr	88057	87998	-	0
	Intr	88219	88166	-	0
20	Intr	88375	88316	-	0
	Intr	88529	88467	-	0
	Intr	88664	88621	-	0
	Intr	88853	88769	-	0
	Intr	89044	88964	-	0
25	Intr	89241	89149	-	0
	Intr	89408	89344	-	0
	Intr	89583	89508	-	0
	Intr	89751	89686	-	0
	Intr	89916	89851	-	0
30	Intr	90281	90192	-	0
	Init	90727	90584	-	0

>2342673 /39667

35	len =	827	nex =	2	
	Init	95406	95717	+	0
	Term	95822	96232	+	0

40 >2342717 /13928

	len =	4710	nex =	16	
	Term	28916	28495	-	0
45	Intr	29102	29002	-	0
	Intr	29276	29211	-	0
	Intr	29479	29365	-	0
	Intr	29760	29654	-	0
	Intr	29937	29848	-	0
50	Intr	30204	30094	-	0
	Intr	30570	30505	-	0
	Intr	30730	30665	-	0
	Intr	31414	31265	-	0
	Intr	31587	31513	-	0
55	Intr	32170	32079	-	0
	Intr	32332	32267	-	0
	Intr	32516	32417	-	0
	Intr	32772	32611	-	0
60	Init	33012	32912	-	0

	>2342717	/23892		
	len =	1550	nex =	4
5	Term	33902	33442	- 0
	Intr	34398	34340	- 0
	Intr	34564	34485	- 0
	Init	34991	34651	- 0
10	>2342717	/25519		
	len =	2805	nex =	5
	Term	38674	38181	- 0
15	Intr	38927	38769	- 0
	Intr	39218	39037	- 0
	Intr	40474	40303	- 0
	Init	40985	40560	- 0
20	>2351061	/36048		
	len =	2257	nex =	4
	Term	36654	36150	- 0
25	Intr	37353	37320	- 0
	Intr	37883	37644	- 0
	Init	38406	38255	- 0
30	>2351061	/16286		
	len =	1302	nex =	2
	Init	60023	60178	+ 0
	Term	60434	60780	+ 0
35	>2351061	/25119		
	len =	2152	nex =	5
40	Init	72312	72460	+ 0
	Intr	72978	73443	+ 0
	Intr	73577	73670	+ 0
	Intr	73763	73893	+ 0
	Term	74106	74463	+ 0
45	>2351061	/7022		
	len =	1348	nex =	1
50	Sngl	74769	74513	- 0
	>2351061	/37512		
	len =	1737	nex =	0
55	>2351062	/1575		
	len =	1492	nex =	2
60	Init	11143	11366	+ 0

				975	
	Term	11952	12270	+	0
	>2351062	/38092			
5	len =	2470	nex =	3	
	Term	27085	26904	-	0
	Intr	28828	27521	-	0
	Init	29365	29247	-	0
10	>2351062	/17241			
	len =	1404	nex =	3	
15	Init	29965	30040	+	0
	Intr	30233	30463	+	0
	Term	30712	30955	+	0
	>2351062	/31041			
20	len =	2710	nex =	8	
	Init	50901	51179	+	0
	Intr	51563	51664	+	0
25	Intr	51779	51832	+	0
	Intr	52010	52102	+	0
	Intr	52264	52356	+	0
	Intr	52687	52791	+	0
	Intr	52881	52979	+	0
30	Term	53072	53603	+	0
	>2351062	/23924			
	len =	1277	nex =	3	
35	Init	71481	71998	+	0
	Intr	72070	72397	+	0
	Term	72483	72757	+	0
40	>2351063	/114691			
	len =	1789	nex =	8	
	Term	20785	20575	-	0
45	Intr	20954	20889	-	0
	Intr	21132	21047	-	0
	Intr	21269	21235	-	0
	Intr	21455	21369	-	0
	Intr	21616	21539	-	0
50	Intr	21741	21701	-	0
	Init	22363	22239	-	0
	>2351063	/36626			
55	len =	1476	nex =	6	
	Term	21132	21053	-	0
	Intr	21269	21235	-	0
	Intr	21455	21369	-	0
60	Intr	21616	21539	-	0

				976	
	Intr	21741	21701	-	0
	Init	22528	22239	-	0
	>2351063	/31913			
5	len =	1211	nex =	4	
	Init	28196	28319	+	0
	Intr	28394	28464	+	0
10	Intr	28552	28573	+	0
	Term	28658	29015	+	0
	>2351063	/103246			
15	len =	1195	nex =	4	
	Init	28196	28319	+	0
	Intr	28394	28464	+	0
	Intr	28552	28573	+	0
20	Term	28658	29015	+	0
	>2351063	/36058			
	len =	2835	nex =	10	
25	Init	55242	55559	+	0
	Intr	55634	55699	+	0
	Intr	55825	55890	+	0
	Intr	56186	56264	+	0
30	Intr	56488	56608	+	0
	Intr	56694	56789	+	0
	Intr	56864	56976	+	0
	Intr	57238	57354	+	0
	Intr	57635	57735	+	0
35	Term	57871	58076	+	0
	>2351063	/95281			
	len =	430	nex =	1	
40	Sngl	58996	59416	+	0
	>2351063	/108981			
45	len =	314	nex =	1	
	Sngl	62819	63132	+	0
	>2351063	/19716			
50	len =	2088	nex =	8	
	Term	66456	66175	-	0
	Intr	66816	66527	-	0
55	Intr	67192	66895	-	0
	Intr	67350	67280	-	0
	Intr	67560	67444	-	0
	Intr	67709	67635	-	0
	Intr	67857	67796	-	0
60	Init	68262	68028	-	0

>2351063 /18140

5	len =	3071	nex =	5	
	Term	81242	80797	-	0
	Intr	81474	81378	-	0
	Intr	81610	81555	-	0
	Intr	81979	81686	-	0
10	Init	82808	82071	-	0

>2351064 /10154

15	len =	1092	nex =	5	
	Init	30526	30610	+	0
	Intr	30871	30941	+	0
	Intr	31032	31188	+	0
	Intr	31364	31450	+	0
20	Term	31536	31617	+	0

>2351064 /23922

25	len =	2156	nex =	9	
	Init	30531	30610	+	0
	Intr	30871	30941	+	0
	Intr	31032	31188	+	0
	Intr	31364	31450	+	0
30	Intr	31536	31687	+	0
	Intr	31802	31882	+	0
	Intr	31983	32091	+	0
	Intr	32233	32359	+	0
35	Term	32454	32686	+	0

>2351064 /41054

	len =	500	nex =	2	
40	Init	32229	32359	+	0
	Term	32454	32728	+	0

>2351064 /37122

45	len =	2271	nex =	5	
	Term	52016	51678	-	0
	Intr	52304	52104	-	0
	Intr	52616	52417	-	0
50	Intr	52811	52698	-	0
	Init	53187	53050	-	0

>2351065 /8508

55	len =	286	nex =	1	
	Sngl	1156	871	-	0

>2351065 /29363

60

978

	len =	8125	nex =	3	
	Term	5274	4953	-	0
	Intr	12650	5804	-	0
5	Init	13070	12743	-	0
	>2351065 /3542				
10	len =	1606	nex =	3	
	Term	12650	12382	-	0
	Intr	13557	12743	-	0
	Init	13987	13679	-	0
15	>2351065 /117588				
	len =	1433	nex =	3	
	Init	26825	26985	+	0
20	Intr	27076	27149	+	0
	Term	27414	28257	+	0
	>2351065 /15229				
25	len =	1952	nex =	9	
	Term	28953	28676	-	0
	Intr	29086	29035	-	0
	Intr	29404	29169	-	0
30	Intr	29662	29605	-	0
	Intr	29821	29753	-	0
	Intr	30022	29914	-	0
	Intr	30232	30165	-	0
	Intr	30434	30315	-	0
35	Init	30627	30561	-	0
	>2351065 /41047				
40	len =	1956	nex =	9	
	Term	28953	28675	-	0
	Intr	29086	29035	-	0
	Intr	29404	29169	-	0
	Intr	29662	29605	-	0
45	Intr	29821	29753	-	0
	Intr	30022	29914	-	0
	Intr	30232	30165	-	0
	Intr	30434	30315	-	0
	Init	30630	30561	-	0
50	>2351065 /105944				
	len =	254	nex =	1	
55	Sngl	38996	38743	-	0
	>2351065 /6823				
60	len =	436	nex =	1	

					979
	Sngl	420	855	+	0
	>2351065	/15640			
5	len =	2139	nex =	7	
	Term	54303	53997	-	0
	Intr	54528	54415	-	0
	Intr	54773	54648	-	0
10	Intr	55027	54948	-	0
	Intr	55198	55117	-	0
	Intr	55390	55316	-	0
	Init	56135	55791	-	0
15	>2351065	/633			
	len =	529	nex =	1	
	Sngl	56522	57050	+	0
20	>2351065	/104017			
	len =	1017	nex =	2	
25	Term	62259	61832	-	0
	Init	62503	62277	-	0
	>2351066	/92216			
30	len =	1063	nex =	1	
	Sngl	2252	1951	-	0
	>2351066	/18332			
35	len =	372	nex =	1	
	Sngl	51067	50696	-	0
40	>2351066	/19255			
	len =	1001	nex =	2	
	Init	6275	6505	+	0
45	Term	6677	6809	+	0
	>2351066	/93148			
	len =	557	nex =	1	
50	Sngl	64963	64407	-	0
	>2351066	/9184			
55	len =	1493	nex =	3	
	Init	65437	65484	+	0
	Intr	65563	65622	+	0
	Term	66328	66800	+	0
60					

	>2351066	/94924			
	len =	772	nex =	4	
5	Term	66989	66757	-	0
	Intr	67176	67069	-	0
	Intr	67314	67274	-	0
	Init	67528	67424	-	0
10	>2351066	/117503			
	len =	1270	nex =	5	
	Term	82968	82813	-	0
15	Intr	83338	83123	-	0
	Intr	83553	83453	-	0
	Intr	83928	83699	-	0
	Init	84064	83998	-	0
20	>2351067	/24137			
	len =	592	nex =	1	
	Sngl	23773	23182	-	0
25	>2351067	/102435			
	len =	1553	nex =	1	
30	Sngl	31407	31589	+	0
	>2351067	/42506			
	len =	913	nex =	2	
35	Init	3624	3761	+	0
	Term	4109	4536	+	0
40	>2351067	/37503			
	len =	1398	nex =	4	
	Term	39519	39286	-	0
	Intr	39736	39638	-	0
45	Intr	40371	40283	-	0
	Init	40683	40599	-	0
	>2351067	/23800			
50	len =	1450	nex =	4	
	Term	39519	39294	-	0
	Intr	39736	39638	-	0
	Intr	40371	40283	-	0
55	Init	40735	40599	-	0
	>2351067	/12458			
60	len =	192	nex =	1	

					981
	Sngl	43705	43896	+	0
	>2351068	/108814			
5	len =	755	nex =	4	
	Init	14299	14392	+	0
	Intr	14508	14644	+	0
	Intr	14771	14817	+	0
10	Term	14906	15053	+	0
	>2351068	/33315			
15	len =	2311	nex =	9	
	Init	14309	14392	+	0
	Intr	14508	14644	+	0
	Intr	14771	14817	+	0
	Intr	14906	15231	+	0
20	Intr	15511	15593	+	0
	Intr	15693	15768	+	0
	Intr	15855	16012	+	0
	Intr	16102	16263	+	0
25	Term	16357	16619	+	0
	>2351068	/37265			
	len =	2272	nex =	9	
30	Init	14347	14392	+	0
	Intr	14508	14644	+	0
	Intr	14771	14817	+	0
	Intr	14906	15231	+	0
	Intr	15511	15593	+	0
35	Intr	15693	15768	+	0
	Intr	15855	16012	+	0
	Intr	16102	16263	+	0
	Term	16357	16618	+	0
40	>2351068	/777			
	len =	540	nex =	1	
	Sngl	22901	23440	+	0
45	>2351068	/2304			
	len =	550	nex =	1	
50	Sngl	22901	23442	+	0
	>2351068	/15211			
	len =	560	nex =	1	
55	Sngl	22904	23463	+	0
	>2351068	/27372			
60	len =	1870	nex =	3	

982

	Init	42505	42957	+	0
	Intr	43205	43414	+	0
	Term	43963	44371	+	0
5	>2351068 /5335				
	len =	731	nex =	3	
10	Init	5218	5304	+	0
	Intr	5320	5477	+	0
	Term	5551	5931	+	0
	>2351068 /22794				
15	len =	857	nex =	1	
	Sngl	61140	61996	+	0
20	>2351068 /28601				
	len =	2200	nex =	7	
	Init	65723	65950	+	0
25	Intr	66035	66198	+	0
	Intr	66298	66349	+	0
	Intr	66544	66771	+	0
	Intr	66874	67063	+	0
	Intr	67153	67418	+	0
30	Term	67680	67922	+	0
	>2351068 /25211				
	len =	1220	nex =	1	
35	Sngl	69965	68746	-	0
	>2351069 /26016				
40	len =	2002	nex =	7	
	Init	26231	26670	+	0
	Intr	26762	26870	+	0
	Intr	26960	27122	+	0
45	Intr	27209	27357	+	0
	Intr	27450	27601	+	0
	Intr	27686	27800	+	0
	Term	27886	28232	+	0
50	>2351069 /1271				
	len =	3480	nex =	9	
	Init	42775	42864	+	0
55	Intr	43235	43369	+	0
	Intr	43517	43633	+	0
	Intr	43791	43942	+	0
	Intr	44014	44098	+	0
	Intr	44277	44371	+	0
60	Intr	44852	45017	+	0

983

	Intr	45150	45345	+	0
	Term	45434	45819	+	0
	>2351069 /13271				
5	len =	702	nex =	3	
	Init	62856	62885	+	0
	Intr	62964	63042	+	0
10	Term	63127	63557	+	0
	>2351069 /7744				
	len =	1618	nex =	8	
15	Term	67305	66948	-	0
	Intr	67508	67411	-	0
	Intr	67723	67598	-	0
	Intr	67896	67813	-	0
20	Intr	68098	67982	-	0
	Intr	68261	68178	-	0
	Intr	68427	68380	-	0
	Init	68562	68508	-	0
25	>2351069 /3285				
	len =	3163	nex =	13	
	Term	67262	67061	-	0
30	Intr	67508	67411	-	0
	Intr	67723	67598	-	0
	Intr	67896	67813	-	0
	Intr	68098	67982	-	0
	Intr	68261	68178	-	0
35	Intr	68427	68380	-	0
	Intr	68562	68508	-	0
	Intr	68759	68704	-	0
	Intr	68928	68844	-	0
	Intr	69102	69029	-	0
40	Intr	69415	69349	-	0
	Init	70098	70008	-	0
	>2351070 /97197				
45	len =	697	nex =	1	
	Sngl	23957	23261	-	0
	>2351070 /6363				
50	len =	560	nex =	1	
	Sngl	34956	34397	-	0
55	>2351070 /26053				
	len =	817	nex =	1	
60	Sngl	46123	46936	+	0

>2351071 /17432

len = 2313 nex = 9

5	Term	46885	46586	-	0
	Intr	47174	47088	-	0
	Intr	47356	47291	-	0
	Intr	47556	47467	-	0
	Intr	47720	47640	-	0
10	Intr	47910	47833	-	0
	Intr	48093	48003	-	0
	Intr	48436	48295	-	0
	Init	48898	48628	-	0

15 >2351071 /39195

len = 2186 nex = 3

	Term	70730	70227	-	0
20	Intr	71606	71158	-	0
	Init	72412	72145	-	0

>2351071 /17360

25 len = 1402 nex = 3

	Term	78193	77927	-	0
	Intr	78535	78274	-	0
	Init	79311	79168	-	0

30 >2351071 /26743

len = 1466 nex = 3

35	Term	78193	77927	-	0
	Intr	78535	78274	-	0
	Init	79392	79168	-	0

>2351072 /29659

40 len = 2508 nex = 5

	Term	22869	22279	-	0
	Intr	23128	23019	-	0
45	Intr	23667	23238	-	0
	Intr	23978	23838	-	0
	Init	24786	24671	-	0

>2351072 /207148

50 len = 797 nex = 1

	Sngl	50991	50195	-	0
--	------	-------	-------	---	---

55 >2351073 /98326

len = 676 nex = 3

60	Term	19588	19334	-	0
	Intr	19757	19681	-	0

					985
	Init	19996	19838	-	0
	>2351073	/100141			
5	len =	1717	nex =	5	
	Term	19588	19293	-	0
	Intr	19757	19681	-	0
	Intr	20220	19838	-	0
10	Intr	20633	20533	-	0
	Init	21009	20902	-	0
	>2351073	/115914			
15	len =	116	nex =	1	
	Sngl	26710	26595	-	0
	>2351073	/95599			
20	len =	749	nex =	3	
	Term	26967	26608	-	0
	Intr	27178	27047	-	0
25	Init	27356	27258	-	0
	>2351073	/35552			
30	len =	1828	nex =	6	
	Term	26967	26653	-	0
	Intr	27178	27047	-	0
	Intr	27399	27258	-	0
	Intr	27742	27550	-	0
35	Intr	28087	27842	-	0
	Init	28480	28170	-	0
	>2351073	/118777			
40	len =	1030	nex =	1	
	Sngl	31871	32900	+	0
	>2358139	/20380			
45	len =	876	nex =	3	
	Init	15794	15936	+	0
	Intr	16035	16176	+	0
50	Term	16428	16669	+	0
	>2358139	/29808			
55	len =	1270	nex =	2	
	Term	64249	63873	-	0
	Init	65100	64760	-	0
	>2358139	/108558			
60					

986

	len =	1069	nex =	3	
	Init	65271	65413	+	0
	Intr	65781	65860	+	0
5	Term	66116	66339	+	0
	>2358139 /1730				
	len =	1484	nex =	3	
10	Init	71725	71848	+	0
	Intr	72291	72590	+	0
	Term	72701	73208	+	0
15	>2392762 /8805				
	len =	1259	nex =	2	
	Term	30586	29909	-	0
20	Init	31167	30868	-	0
	>2392762 /14724				
	len =	1796	nex =	8	
25	Term	60877	60621	-	0
	Intr	61051	60973	-	0
	Intr	61293	61140	-	0
	Intr	61514	61420	-	0
30	Intr	61620	61585	-	0
	Intr	61952	61727	-	0
	Intr	62107	62037	-	0
	Init	62416	62342	-	0
35	>2392762 /15990				
	len =	1729	nex =	8	
	Term	60877	60688	-	0
40	Intr	61051	60973	-	0
	Intr	61293	61140	-	0
	Intr	61514	61420	-	0
	Intr	61620	61585	-	0
	Intr	61952	61727	-	0
45	Intr	62107	62037	-	0
	Init	62416	62342	-	0
	>2392762 /41162				
50	len =	951	nex =	3	
	Init	68249	68350	+	0
	Intr	68449	68513	+	0
	Term	68901	69199	+	0
55	>2435510 /32833				
	len =	1450	nex =	5	
60	Term	41015	40654	-	0

					987
	Intr	41265	41098	-	0
	Intr	41451	41368	-	0
	Intr	41718	41540	-	0
	Init	42097	41892	-	0
5	>2435510	/1011			
	len =	1120	nex =	3	
10	Term	51801	51490	-	0
	Intr	52028	51949	-	0
	Init	52609	52122	-	0
	>2435510	/19362			
15	len =	1041	nex =	3	
	Init	61031	61254	+	0
	Intr	61359	61535	+	0
20	Term	61610	62071	+	0
	>2435510	/142314			
	len =	919	nex =	3	
25	Init	61151	61254	+	0
	Intr	61359	61535	+	0
	Term	61610	62069	+	0
	>2435510	/33456			
30	len =	2142	nex =	6	
	Term	4364	4051	-	0
35	Intr	4676	4612	-	0
	Intr	5214	5151	-	0
	Intr	5423	5314	-	0
	Intr	5600	5513	-	0
	Init	6192	5794	-	0
40	>2435510	/4367			
	len =	2039	nex =	8	
45	Init	76018	76119	+	0
	Intr	76377	76574	+	0
	Intr	76648	76707	+	0
	Intr	76793	77235	+	0
	Intr	77335	77501	+	0
50	Intr	77587	77660	+	0
	Intr	77749	77808	+	0
	Term	77912	78053	+	0
	>2443899	/22008			
55	len =	1489	nex =	2	
	Term	102074	101797	-	0
	Init	103282	102296	-	0
60					

>2443899 /1734

len = 888 nex = 2

5	Term	14747	14318	-	0
	Init	15205	14965	-	0

>2459406 /42992

10 len = 2396 nex = 10

	Term	117911	117825	-	0
	Intr	118071	117986	-	0
	Intr	118340	118166	-	0
15	Intr	118518	118458	-	0
	Intr	118661	118595	-	0
	Intr	118838	118754	-	0
	Intr	119077	118920	-	0
	Intr	119310	119166	-	0
20	Intr	119486	119427	-	0
	Init	119855	119575	-	0

>2459406 /11254

25 len = 2035 nex = 6

	Init	128392	128598	+	0
	Intr	128894	129063	+	0
	Intr	129142	129327	+	0
30	Intr	129412	129577	+	0
	Intr	129681	129870	+	0
	Term	130089	130426	+	0

>2459406 /92741

35 len = 538 nex = 1

	Sngl	141230	140693	-	0
--	------	--------	--------	---	---

40 >2459406 /13741

len = 1713 nex = 4

	Term	18475	18146	-	0
45	Intr	18628	18567	-	0
	Intr	19123	18713	-	0
	Init	19858	19394	-	0

>2459406 /25272

50 len = 1750 nex = 4

	Init	2679	2985	+	0
	Intr	3377	3419	+	0
55	Intr	3511	3571	+	0
	Term	3697	4419	+	0

>2459406 /35273

60 len = 2218 nex = 3

989

	Term	26889	26777	-	0
	Intr	28208	27837	-	0
	Init	28994	28459	-	0
5	>2459406 /28563				
	len =	1150	nex =	4	
10	Term	47656	47428	-	0
	Intr	47792	47751	-	0
	Intr	48158	47874	-	0
	Init	48577	48488	-	0
15	>2459406 /119409				
	len =	468	nex =	1	
	Sngl	57470	57023	-	0
20	>2459406 /116034				
	len =	337	nex =	1	
25	Sngl	61222	61558	+	0
	>2459406 /8717				
	len =	2113	nex =	4	
30	Init	66546	66940	+	0
	Intr	67084	67181	+	0
	Intr	67274	67339	+	0
	Term	68443	68658	+	0
35	>2459406 /31633				
	len =	945	nex =	2	
40	Init	77435	77674	+	0
	Term	78004	78379	+	0
	>2459406 /19302				
45	len =	2115	nex =	6	
	Term	80490	80306	-	0
	Intr	80717	80586	-	0
	Intr	80949	80814	-	0
50	Intr	81174	81044	-	0
	Intr	81479	81424	-	0
	Init	82420	82270	-	0
	>2459406 /37919				
55	len =	2274	nex =	6	
	Term	80490	80262	-	0
	Intr	80717	80586	-	0
60	Intr	80949	80814	-	0

				990	
	Intr	81174	81044	-	0
	Intr	81479	81424	-	0
	Init	82535	82270	-	0
5	>2459406	/18894			
	len =	235	nex =	1	
	Sngl	85070	85304	+	0
10	>2477521	/15308			
	len =	1434	nex =	1	
15	Sngl	11192	12625	+	0
	>2477521	/27205			
	len =	760	nex =	3	
20	Term	22663	22447	-	0
	Intr	22864	22743	-	0
	Init	23206	22955	-	0
25	>2477521	/40049			
	len =	3210	nex =	5	
	Init	52491	52536	+	0
30	Intr	52618	52732	+	0
	Intr	52824	52891	+	0
	Intr	52986	53708	+	0
	Term	53792	54336	+	0
35	>2477521	/3549			
	len =	1750	nex =	4	
	Init	59783	60056	+	0
40	Intr	60329	60677	+	0
	Intr	60773	60914	+	0
	Term	60979	61527	+	0
	>2477521	/12293			
45	len =	1796	nex =	6	
	Term	71123	70636	-	0
	Intr	71380	71205	-	0
50	Intr	71502	71478	-	0
	Intr	71702	71620	-	0
	Intr	72024	71951	-	0
	Init	72431	72108	-	0
55	>2477521	/98850			
	len =	4463	nex =	7	
	Init	74583	74814	+	0
60	Intr	77407	77441	+	0

```

                                     991
      Intr  77553  77614      +      0
      Intr  77696  77795      +      0
      Intr  77904  77945      +      0
      Intr  78281  78322      +      0
5      Term  78695  79045      +      0

>2477521      /92459

      len =    4460  nex =      7
10      Init  74588  74814      +      0
      Intr  77285  77342      +      0
      Intr  77553  77614      +      0
      Intr  77696  77795      +      0
15      Intr  77904  77945      +      0
      Intr  78281  78322      +      0
      Term  78695  79047      +      0

>2477521      /5076
20      len =     730  nex =      3

      Term  79591  79372      -      0
      Intr  79924  79697      -      0
25      Init  80096  80042      -      0

>2477521      /4033

      len =    1930  nex =      7
30      Init  94403  94493      +      0
      Intr  94625  94761      +      0
      Intr  94865  94911      +      0
      Intr  94999  95483      +      0
35      Intr  95570  95727      +      0
      Intr  95814  95975      +      0
      Term  96051  96327      +      0

>2494106      /36412
40      len =    1375  nex =      3

      Term  99606  98923      -      0
      Intr 100124  99692      -      0
45      Init 100297 100214      -      0

>2494106      /11408

      len =     644  nex =      1
50      Sngl 109531 110174      +      0

>2494106      /8951

55      len =     910  nex =      1

      Sngl 112974 112773      -      0

>2494106      /37020
60

```

992

	len =	757	nex =	4	
	Term	122980	122712	-	0
	Intr	123133	123078	-	0
5	Intr	123278	123220	-	0
	Init	123468	123370	-	0
	>2494106 /29872				
10	len =	861	nex =	2	
	Term	122980	122712	-	0
	Init	123133	123078	-	0
15	>2494106 /34434				
	len =	866	nex =	4	
	Term	122980	122714	-	0
20	Intr	123133	123078	-	0
	Intr	123278	123220	-	0
	Init	123577	123370	-	0
	>2494106 /34374				
25	len =	359	nex =	2	
	Term	123278	123219	-	0
	Init	123577	123370	-	0
30	>2494106 /5465				
	len =	2050	nex =	7	
35	Init	132597	132734	+	0
	Intr	133129	133207	+	0
	Intr	133336	133389	+	0
	Intr	133680	133793	+	0
	Intr	134040	134107	+	0
40	Intr	134190	134301	+	0
	Term	134381	134640	+	0
	>2494106 /1520				
45	len =	1810	nex =	6	
	Init	132677	133207	+	0
	Intr	133336	133389	+	0
	Intr	133680	133793	+	0
50	Intr	134040	134107	+	0
	Intr	134190	134301	+	0
	Term	134381	134477	+	0
	>2494106 /2681				
55	len =	910	nex =	1	
	Sngl	143514	143911	+	0
60	>2494106 /33770				

	len =	1302	nex =	4	
	Term	158712	158351	-	0
5	Intr	159059	158976	-	0
	Intr	159236	159156	-	0
	Init	159509	159332	-	0
	>2494106 /27457				
10	len =	1462	nex =	4	
	Term	40898	40547	-	0
	Intr	41137	41003	-	0
15	Intr	41443	41231	-	0
	Init	42008	41526	-	0
	>2494106 /25255				
20	len =	1719	nex =	3	
	Init	54004	54063	+	0
	Intr	54151	54486	+	0
	Term	54639	54877	+	0
25	>2494106 /14939				
	len =	610	nex =	2	
30	Init	54277	54486	+	0
	Term	54639	54879	+	0
	>2494106 /32130				
35	len =	3130	nex =	12	
	Term	56042	55686	-	0
	Intr	56181	56114	-	0
	Intr	56328	56265	-	0
40	Intr	56502	56421	-	0
	Intr	56676	56618	-	0
	Intr	56984	56925	-	0
	Intr	57266	57104	-	0
	Intr	57498	57374	-	0
45	Intr	57857	57795	-	0
	Intr	58060	58001	-	0
	Intr	58325	58140	-	0
	Init	58811	58689	-	0
50	>2494106 /6667				
	len =	1554	nex =	1	
	Sngl	60644	59091	-	0
55	>2494106 /25894				
	len =	1630	nex =	4	
60	Term	64139	63599	-	0

				994	
	Intr	64439	64381	-	0
	Intr	64965	64855	-	0
	Init	65226	65138	-	0
5	>2494110	/23300			
	len =	2036	nex =	5	
	Term	17775	17469	-	0
10	Intr	18041	17877	-	0
	Intr	18302	18159	-	0
	Intr	18618	18423	-	0
	Init	19504	19053	-	0
15	>2494110	/8559			
	len =	1302	nex =	2	
	Init	25200	25402	+	0
20	Term	26210	26501	+	0
	>2494110	/37952			
	len =	4214	nex =	6	
25	Init	25200	25402	+	0
	Intr	26210	26290	+	0
	Intr	27617	28259	+	0
	Intr	28358	28461	+	0
30	Intr	28571	28709	+	0
	Term	28803	29413	+	0
	>2494110	/21100			
35	len =	812	nex =	3	
	Term	30699	30410	-	0
	Intr	30921	30796	-	0
	Init	31221	30993	-	0
40	>2494110	/34753			
	len =	807	nex =	3	
45	Term	30699	30415	-	0
	Intr	30921	30796	-	0
	Init	31221	30993	-	0
	>2494110	/110726			
50	len =	493	nex =	1	
	Sngl	32194	32672	+	0
55	>2494110	/2265			
	len =	494	nex =	1	
60	Sngl	38819	39312	+	0

995

	>2494110	/13232		
	len =	1220	nex =	1
5	Sngl	40544	39752	- 0
	>2494110	/31923		
	len =	1284	nex =	3
10	Init	41985	42310	+ 0
	Intr	42859	42930	+ 0
	Term	43017	43268	+ 0
15	>2494110	/100984		
	len =	1340	No match - No prediction	
	>2494110	/27110		
20	len =	108	nex =	1
	Sngl	74373	74480	+ 0
	>2494110	/40608		
25	len =	1703	nex =	5
	Term	91321	90966	- 0
	Intr	91466	91405	- 0
30	Intr	91657	91540	- 0
	Intr	92025	91739	- 0
	Init	92668	92298	- 0
	>2494110	/2935		
35	len =	1613	nex =	5
	Init	97175	97627	+ 0
	Intr	97725	97897	+ 0
40	Intr	97974	98088	+ 0
	Intr	98324	98478	+ 0
	Term	98578	98787	+ 0
	>2505864	/35333		
45	len =	1426	nex =	4
	Init	20951	21020	+ 0
	Intr	21255	21415	+ 0
50	Intr	21681	21869	+ 0
	Term	22136	22367	+ 0
	>2505864	/4328		
55	len =	1374	nex =	4
	Init	21013	21070	+ 0
	Intr	21255	21415	+ 0
	Intr	21681	21869	+ 0
60	Term	22136	22386	+ 0


```

>2505873      /6115
5      len =    1600    nex =    2
      Term  14088    13696    -    0
      Init  15295    14419    -    0

>2505873      /33852
10     len =    134    nex =    1
      Sngl  19922    20055    +    0

15     >2505873      /36699
      len =    236    nex =    1
      Sngl  27483    27718    +    0

20     >2529657      /32457
      len =    1232    nex =    5
25     Term  10325    10206    -    0
      Intr  10512    10408    -    0
      Intr  10777    10703    -    0
      Intr  11135    11111    -    0
      Init  11437    11243    -    0

30     >2529657      /26123
      len =    1422    nex =    5
35     Term  10325    10041    -    0
      Intr  10512    10408    -    0
      Intr  10777    10703    -    0
      Intr  11135    11111    -    0
      Init  11462    11380    -    0

40     >2529657      /20647
      len =    1390    nex =    5
45     Term  12109    11630    -    0
      Intr  12283    12185    -    0
      Intr  12499    12362    -    0
      Intr  12722    12592    -    0
      Init  13015    12840    -    0

50     >2529657      /28691
      len =    2057    nex =    7
55     Term  12109    11913    -    0
      Intr  12283    12185    -    0
      Intr  12499    12362    -    0
      Intr  12722    12592    -    0
      Intr  12986    12840    -    0
60     Intr  13647    13615    -    0

```

997

	Init	13969	13735	-	0
	>2529657	/33373			
5	len =	2492	nex =	8	
	Term	12109	11637	-	0
	Intr	12283	12185	-	0
	Intr	12499	12362	-	0
10	Intr	12722	12592	-	0
	Intr	12986	12840	-	0
	Intr	13647	13615	-	0
	Intr	13831	13735	-	0
	Init	14128	13974	-	0
15	>2529657	/24272			
	len =	1054	nex =	4	
20	Term	17370	17243	-	0
	Intr	17555	17463	-	0
	Intr	17935	17637	-	0
	Init	18296	18094	-	0
25	>2529657	/6394			
	len =	1870	nex =	5	
	Term	17370	16988	-	0
30	Intr	17555	17463	-	0
	Intr	17935	17637	-	0
	Intr	18295	18094	-	0
	Init	18459	18415	-	0
35	>2529657	/25729			
	len =	802	nex =	1	
	Sngl	3834	4635	+	0
40	>2529657	/37870			
	len =	3805	nex =	17	
45	Term	46706	46424	-	0
	Intr	46947	46882	-	0
	Intr	47087	47058	-	0
	Intr	47280	47182	-	0
	Intr	47466	47371	-	0
50	Intr	47623	47573	-	0
	Intr	47773	47707	-	0
	Intr	47950	47856	-	0
	Intr	48158	48077	-	0
	Intr	48324	48275	-	0
55	Intr	48463	48413	-	0
	Intr	48638	48540	-	0
	Intr	49052	48969	-	0
	Intr	49302	49192	-	0
	Intr	49575	49426	-	0
60	Intr	49795	49678	-	0

Reference No. 2750-942P

					998
	Init	50050	49884	-	0
	>2529657	/32039			
5	len =	670	nex =	1	
	Sngl	63987	64654	+	0
	>2529657	/9499			
10	len =	654	nex =	2	
	Term	65131	64658	-	0
	Init	65297	65222	-	0
15	>2529657	/38461			
	len =	2350	nex =	10	
20	Term	65131	64823	-	0
	Intr	65346	65222	-	0
	Intr	65588	65432	-	0
	Intr	65777	65686	-	0
	Intr	65890	65863	-	0
25	Intr	66093	65976	-	0
	Intr	66394	66339	-	0
	Intr	66604	66507	-	0
	Intr	66777	66693	-	0
	Init	67165	66986	-	0
30	>2529657	/13774			
	len =	2397	nex =	10	
35	Term	65131	64823	-	0
	Intr	65346	65222	-	0
	Intr	65588	65432	-	0
	Intr	65777	65686	-	0
	Intr	65890	65863	-	0
40	Intr	66093	65976	-	0
	Intr	66394	66339	-	0
	Intr	66604	66507	-	0
	Intr	66777	66693	-	0
	Init	67219	66986	-	0
45	>2529657	/34914			
	len =	717	nex =	1	
50	Sngl	75255	74539	-	0
	>2529657	/37980			
	len =	1352	nex =	1	
55	Sngl	75893	74542	-	0
	>2564044	/156017			
60	len =	401	nex =	1	

	Sngl	12975	12575	-	0
5	>2564044	/156773			
	len =	350	nex =	1	
	Sngl	12997	12648	-	0
10	>2564044	/31129			
	len =	430	nex =	1	
15	Sngl	13041	12616	-	0
	>2564044	/21629			
	len =	1610	nex =	5	
20	Term	36986	36739	-	0
	Intr	37123	37068	-	0
	Intr	37318	37272	-	0
	Intr	37669	37626	-	0
	Init	38348	38232	-	0
25	>2564044	/22860			
	len =	3400	nex =	11	
30	Init	5043	5315	+	0
	Intr	5670	5734	+	0
	Intr	5871	5969	+	0
	Intr	6171	6303	+	0
	Intr	6748	6807	+	0
35	Intr	6897	7019	+	0
	Intr	7379	7450	+	0
	Intr	7562	7699	+	0
	Intr	7786	7941	+	0
	Intr	8028	8132	+	0
40	Term	8282	8442	+	0
	>2564045	/108335			
45	len =	1516	nex =	2	
	Term	653	118	-	0
	Init	1633	770	-	0
50	>2564045	/512			
	len =	1435	nex =	1	
	Sngl	40196	39668	-	0
55	>2564045	/40250			
	len =	1210	nex =	2	
60	Term	57008	56234	-	0
	Init	57441	57096	-	0

1000

	>2564045	/36090			
5	len =	1219	nex =	2	
	Term	57008	56234	-	0
	Init	57452	57096	-	0
10	>2564045	/33763			
	len =	1217	nex =	1	
	Sngl	5886	7102	+	0
15	>2564045	/23566			
	len =	1043	nex =	3	
	Init	9042	9192	+	0
20	Intr	9618	9763	+	0
	Term	9851	10084	+	0
	>2564046	/4272			
25	len =	4185	nex =	11	
	Term	18249	17894	-	0
	Intr	18506	18454	-	0
	Intr	18683	18598	-	0
30	Intr	18985	18867	-	0
	Intr	19502	19431	-	0
	Intr	19881	19708	-	0
	Intr	20444	20289	-	0
	Intr	20917	20836	-	0
35	Intr	21276	21130	-	0
	Intr	21654	21468	-	0
	Init	22078	21842	-	0
	>2564046	/13993			
40	len =	1672	nex =	6	
	Init	27089	27339	+	0
	Intr	27573	27725	+	0
45	Intr	27820	27972	+	0
	Intr	28179	28262	+	0
	Intr	28344	28485	+	0
	Term	28581	28760	+	0
50	>2564046	/35683			
	len =	697	nex =	3	
	Term	34417	34208	-	0
55	Intr	34609	34504	-	0
	Init	34904	34742	-	0
	>2564047	/12802			
60	len =	1648	nex =	3	

1001

	Init	16402	16518	+	0
	Intr	17081	17129	+	0
	Term	17663	17714	+	0
5	>2564047 /19442				
	len =	850	nex =	1	
10	Sngl	21861	22701	+	0
	>2564047 /2533				
	len =	1779	nex =	3	
15	Init	37480	37886	+	0
	Intr	37970	38637	+	0
	Term	39199	39258	+	0
20	>2564047 /32890				
	len =	1279	nex =	1	
	Sngl	51389	50111	-	0
25	>2564047 /13737				
	len =	1302	nex =	5	
30	Term	57880	57668	-	0
	Intr	58070	58011	-	0
	Intr	58297	58197	-	0
	Intr	58633	58398	-	0
	Init	58969	58725	-	0
35	>2564047 /6893				
	len =	1309	nex =	5	
40	Term	57880	57662	-	0
	Intr	58070	58011	-	0
	Intr	58297	58197	-	0
	Intr	58633	58398	-	0
	Init	58970	58725	-	0
45	>2564047 /114864				
	len =	2470	nex =	9	
50	Init	59318	59464	+	0
	Intr	59652	59723	+	0
	Intr	59821	59895	+	0
	Intr	60508	60588	+	0
	Intr	60854	60923	+	0
55	Intr	60996	61087	+	0
	Intr	61178	61219	+	0
	Intr	61298	61378	+	0
	Term	61566	61785	+	0
60	>2564047 /105566				

1002

	len =	979	nex =	1	
5	Sngl	62070	63048	+	0
	>2564047	/12455			
	len =	1933	nex =	2	
10	Term	67046	66415	-	0
	Init	68347	67916	-	0
	>2564047	/40711			
15	len =	850	nex =	1	
	Sngl	78369	77529	-	0
	>2564048	/105906			
20	len =	1212	nex =	2	
	Init	2380	2769	+	0
	Term	2946	3591	+	0
25	>2564048	/115613			
	len =	586	nex =	1	
30	Sngl	31514	30929	-	0
	>2564048	/1200			
35	len =	1778	nex =	4	
	Term	38609	37885	-	0
	Intr	38864	38681	-	0
	Intr	39244	38988	-	0
	Init	39662	39331	-	0
40	>2564048	/39462			
	len =	2351	nex =	7	
45	Init	41518	41825	+	0
	Intr	42059	42268	+	0
	Intr	42387	42557	+	0
	Intr	42766	42889	+	0
	Intr	43155	43216	+	0
50	Intr	43305	43386	+	0
	Term	43481	43868	+	0
	>2564048	/10292			
55	len =	1951	nex =	7	
	Init	41667	41825	+	0
	Intr	42059	42268	+	0
	Intr	42387	42557	+	0
60	Intr	42766	42889	+	0

1003

Intr	43155	43216	+	0
Intr	43305	43386	+	0
Term	43481	43617	+	0

5 >2564048 /26637

len = 1980 nex = 7

10	Init	61938	62027	+	0
	Intr	62306	62497	+	0
	Intr	62586	62757	+	0
	Intr	62859	62932	+	0
	Intr	63011	63037	+	0
	Intr	63126	63149	+	0
15	Term	63245	63657	+	0

>2564048 /158431

20 len = 1690 nex = 5

	Init	65258	65519	+	0
	Intr	65699	65751	+	0
	Intr	65845	65980	+	0
	Intr	66115	66290	+	0
25	Term	66365	66942	+	0

>2564049 /37294

30 len = 1427 nex = 2

	Term	485	294	-	0
	Init	1720	625	-	0

>2564049 /104793

35 len = 1873 nex = 7

	Init	17973	18128	+	0
	Intr	18663	18789	+	0
40	Intr	18882	19035	+	0
	Intr	19112	19208	+	0
	Intr	19304	19392	+	0
	Intr	19521	19589	+	0
45	Term	19790	19845	+	0

>2564049 /141731

len = 2068 nex = 7

50	Init	18007	18128	+	0
	Intr	18663	18789	+	0
	Intr	18882	19035	+	0
	Intr	19112	19208	+	0
	Intr	19304	19392	+	0
55	Intr	19521	19589	+	0
	Term	19790	20074	+	0

>2564049 /21604

60 len = 557 nex = 1

1004

	Sngl	28618	28062	-	0
	>2564049	/16144			
5	len =	1365	nex =	3	
	Init	28919	29348	+	0
	Intr	29603	29695	+	0
10	Term	30029	30283	+	0
	>2564049	/31971			
	len =	1818	nex =	2	
15	Init	35677	36089	+	0
	Term	36890	37494	+	0
	>2564049	/13667			
20	len =	1704	nex =	6	
	Term	5026	4812	-	0
	Intr	5207	5118	-	0
25	Intr	5466	5299	-	0
	Intr	5691	5572	-	0
	Intr	5932	5787	-	0
	Init	6515	6354	-	0
30	>2564050	/6203			
	len =	2839	nex =	13	
	Init	12017	12391	+	0
35	Intr	12485	12567	+	0
	Intr	12820	12974	+	0
	Intr	13048	13082	+	0
	Intr	13144	13293	+	0
	Intr	13467	13562	+	0
40	Intr	13634	13750	+	0
	Intr	13832	13951	+	0
	Intr	14029	14121	+	0
	Intr	14202	14324	+	0
	Intr	14407	14523	+	0
45	Intr	14606	14668	+	0
	Term	14766	14842	+	0
	>2564050	/123496			
50	len =	674	nex =	1	
	Sngl	17696	18369	+	0
	>2564050	/16313			
55	len =	1594	nex =	5	
	Init	2671	2918	+	0
	Intr	3227	3325	+	0
60	Intr	3410	3518	+	0

					1005
	Intr	3687	3758	+	0
	Term	3993	4264	+	0
5	>2564050	/14738			
	len =	1040	nex =	2	
	Term	28103	27853	-	0
10	Init	28892	28606	-	0
	>2564050	/13951			
	len =	1063	nex =	2	
15	Term	28103	27848	-	0
	Init	28910	28606	-	0
	>2564050	/38057			
20	len =	2006	nex =	0	
	>2564051	/7688			
	len =	1722	nex =	2	
25	Term	13311	12928	-	0
	Init	13996	13887	-	0
	>2564051	/6220			
30	len =	2570	nex =	6	
	Init	18254	18493	+	0
	Intr	18575	18754	+	0
35	Intr	19785	19904	+	0
	Intr	19917	20078	+	0
	Intr	20178	20459	+	0
	Term	20546	20823	+	0
40	>2564051	/30648			
	len =	2334	nex =	9	
	Init	33401	33589	+	0
45	Intr	33676	33848	+	0
	Intr	34149	34268	+	0
	Intr	34373	34429	+	0
	Intr	34595	34675	+	0
	Intr	34763	34797	+	0
50	Intr	34933	35006	+	0
	Intr	35103	35262	+	0
	Term	35380	35734	+	0
	>2564051	/30994			
55	len =	2530	nex =	8	
	Init	45513	45608	+	0
	Intr	46036	46115	+	0
60	Intr	46206	46280	+	0

Reference No. 2750-942P

				1006	
	Intr	46370	46473	+	0
	Intr	46561	46717	+	0
	Intr	46810	46897	+	0
	Intr	46997	47069	+	0
5	Term	47147	47224	+	0
	>2564051		/29619		
10	len =	942	nex =	3	
	Init	46810	46897	+	0
	Intr	46997	47069	+	0
	Term	47147	47224	+	0
15	>2564051		/29829		
	len =	1317	nex =	3	
20	Term	48114	47710	-	0
	Intr	48493	48207	-	0
	Init	49026	48809	-	0
	>2564051		/6519		
25	len =	1128	nex =	2	
	Init	72721	72978	+	0
	Term	73194	73848	+	0
30	>2564051		/142033		
	len =	651	nex =	2	
35	Init	72788	72978	+	0
	Term	73194	73438	+	0
	>2564051		/14159		
40	len =	1394	nex =	5	
	Term	74311	74056	-	0
	Intr	74603	74398	-	0
	Intr	74863	74713	-	0
	Intr	75172	74950	-	0
45	Init	75449	75412	-	0
	>2564051		/40866		
50	len =	1519	nex =	5	
	Term	74311	74064	-	0
	Intr	74603	74398	-	0
	Intr	74863	74713	-	0
	Intr	75172	74950	-	0
55	Init	75582	75412	-	0
	>2564051		/17770		
60	len =	1500	nex =	5	

Reference No. 2750-942P

				1007	
	Term	74311	74086	-	0
	Intr	74603	74398	-	0
	Intr	74863	74713	-	0
	Intr	75172	74950	-	0
5	Init	75445	75412	-	0
	>2564051 /13949				
10	len =	1110	nex =	3	
	Term	82879	82476	-	0
	Intr	83240	82973	-	0
	Init	83585	83325	-	0
15	>2570223 /40832				
	len =	2253	nex =	6	
20	Init	17162	17477	+	0
	Intr	17799	17892	+	0
	Intr	18430	18609	+	0
	Intr	18688	18807	+	0
	Intr	18887	19020	+	0
25	Term	19185	19414	+	0
	>2570223 /37699				
	len =	2869	nex =	9	
30	Term	26477	25979	-	0
	Intr	26840	26580	-	0
	Intr	27159	26941	-	0
	Intr	27498	27271	-	0
	Intr	27878	27776	-	0
35	Intr	28077	27965	-	0
	Intr	28258	28197	-	0
	Intr	28478	28346	-	0
	Init	28847	28757	-	0
40	>2570223 /23106				
	len =	629	nex =	1	
45	Sngl	74691	75319	+	0
	>2583106 /29207				
	len =	2272	nex =	5	
50	Init	108141	108430	+	0
	Intr	108875	109071	+	0
	Intr	109540	109629	+	0
	Intr	109744	109815	+	0
55	Term	110152	110412	+	0
	>2583106 /36389				
	len =	643	nex =	1	
60	Sngl	121587	120945	-	0

1008

	>2583106	/17187			
5	len =	677	nex =	1	
	Sngl	121618	120942	-	0
	>2583106	/23203			
10	len =	704	nex =	2	
	Init	13956	14106	+	0
	Term	14207	14659	+	0
15	>2583106	/2322			
	len =	531	nex =	1	
20	Sngl	15480	14950	-	0
	>2583106	/26817			
	len =	1955	nex =	3	
25	Term	15998	14950	-	0
	Intr	16202	16119	-	0
	Init	16904	16559	-	0
30	>2583106	/7709			
	len =	1471	nex =	1	
	Sngl	3827	5297	+	0
35	>2583106	/33864			
	len =	1700	nex =	1	
40	Sngl	64128	64474	+	0
	>2583106	/27799			
	len =	1690	nex =	4	
45	Term	73308	72358	-	0
	Intr	73553	73400	-	0
	Intr	73796	73648	-	0
	Init	74047	73886	-	0
50	>2583106	/15659			
	len =	2848	nex =	7	
55	Term	73308	72358	-	0
	Intr	73553	73400	-	0
	Intr	73796	73648	-	0
	Intr	74168	73886	-	0
	Intr	74356	74251	-	0
	Intr	74536	74446	-	0
60	Init	75205	74719	-	0

1009

	>2583106	/18320			
5	len =	2568	nex =	7	
	Term	73308	72638	-	0
	Intr	73553	73400	-	0
	Intr	73796	73648	-	0
	Intr	74168	73886	-	0
10	Intr	74356	74251	-	0
	Intr	74536	74446	-	0
	Init	75205	74719	-	0
	>2583106	/21765			
15	len =	1874	nex =	0	
	>2583106	/1969			
20	len =	5689	nex =	1	
	Sngl	88355	88684	+	0
	>2583106	/37127			
25	len =	310	nex =	0	
	>2583106	/37621			
30	len =	3101	nex =	15	
	Term	89198	88862	-	0
	Intr	89371	89306	-	0
	Intr	89531	89462	-	0
35	Intr	89689	89616	-	0
	Intr	89891	89793	-	0
	Intr	90037	89976	-	0
	Intr	90178	90137	-	0
	Intr	90316	90265	-	0
40	Intr	90541	90442	-	0
	Intr	90682	90638	-	0
	Intr	90843	90796	-	0
	Intr	91179	91104	-	0
	Intr	91456	91286	-	0
45	Intr	91590	91540	-	0
	Init	91962	91806	-	0
	>2584827	/273			
50	len =	2260	nex =	5	
	Term	101872	101586	-	0
	Intr	102093	102017	-	0
	Intr	102388	102242	-	0
55	Intr	102650	102480	-	0
	Init	103150	102928	-	0
	>2584827	/5480			
60	len =	1994	nex =	3	

1010

	Term	111925	111310	-	0
	Intr	112167	112049	-	0
	Init	113303	112263	-	0
5	>2584827 /5171				
	len =	319	nex =	1	
10	Sngl	115114	114796	-	0
	>2584827 /17426				
	len =	597	nex =	1	
15	Sngl	115422	114826	-	0
	>2584827 /11593				
20	len =	562	nex =	1	
	Sngl	115422	114861	-	0
	>2584827 /25571				
25	len =	610	nex =	1	
	Sngl	115430	114821	-	0
30	>2584827 /34348				
	len =	1756	nex =	8	
	Term	117147	116843	-	0
35	Intr	117385	117233	-	0
	Intr	117590	117483	-	0
	Intr	117734	117687	-	0
	Intr	118025	117813	-	0
	Intr	118181	118117	-	0
40	Intr	118386	118262	-	0
	Init	118595	118482	-	0
	>2584827 /39107				
45	len =	2383	nex =	7	
	Term	117147	116840	-	0
	Intr	117385	117233	-	0
	Intr	117590	117483	-	0
50	Intr	117734	117687	-	0
	Intr	118025	117813	-	0
	Intr	118181	118117	-	0
	Init	118386	118262	-	0
55	>2584827 /5712				
	len =	790	nex =	1	
60	Sngl	23900	24682	+	0

1011

	>2584827	/27675			
	len =	745	nex =	1	
5	Sngl	23978	24722	+	0
	>2584827	/116395			
10	len =	286	nex =	1	
	Sngl	29868	29583	-	0
	>2584827	/4503			
15	len =	471	nex =	1	
	Sngl	30113	29649	-	0
20	>2584827	/22292			
	len =	592	nex =	1	
	Sngl	30233	29642	-	0
25	>2584827	/25064			
	len =	683	nex =	2	
30	Term	30138	29660	-	0
	Init	30342	30322	-	0
	>2584827	/25142			
35	len =	816	nex =	2	
	Term	30138	29596	-	0
	Init	30411	30322	-	0
40	>2584827	/1994			
	len =	835	nex =	2	
	Term	30138	29584	-	0
45	Init	30418	30322	-	0
	>2584827	/4479			
	len =	655	nex =	3	
50	Term	84398	84092	-	0
	Intr	84584	84486	-	0
	Init	84746	84674	-	0
55	>2584827	/31676			
	len =	649	nex =	2	
	Term	85483	85211	-	0
60	Init	85859	85582	-	0

1012

>2584827

/32472

len = 1762 nex = 7

5	Term	84398	84113	-	0
	Intr	84584	84486	-	0
	Intr	84807	84674	-	0
	Intr	84997	84910	-	0
	Intr	85301	85255	-	0
10	Intr	85483	85417	-	0
	Init	85870	85582	-	0

>2584827

/8972

15	len =	4044	nex =	12	
	Init	95183	95243	+	0
	Intr	95429	95523	+	0
	Intr	95608	95720	+	0
20	Intr	95804	95972	+	0
	Intr	96059	96098	+	0
	Intr	96231	96295	+	0
	Intr	96387	96500	+	0
	Intr	96601	96665	+	0
25	Intr	96783	96939	+	0
	Intr	97037	97156	+	0
	Intr	97247	97335	+	0
	Term	97422	97750	+	0

30 >2584827 /17473

len = 1881 nex = 9

	Init	95871	95972	+	0
35	Intr	96059	96098	+	0
	Intr	96231	96295	+	0
	Intr	96387	96500	+	0
	Intr	96601	96665	+	0
	Intr	96783	96939	+	0
40	Intr	97037	97156	+	0
	Intr	97247	97335	+	0
	Term	97422	97751	+	0

>2618599 /23293

45	len =	776	nex =	1	
	Sngl	11508	11343	-	0

50 >2618599 /6500

len = 997 nex = 1

	Sngl	13043	14039	+	0
--	------	-------	-------	---	---

>2618599 /40212

len = 1750 nex = 3

60	Term	12611	12066	-	0
----	------	-------	-------	---	---

					1013
	Intr	13271	13086	-	0
	Init	13808	13354	-	0
	>2618599	/39514			
5	len =	1757	nex =	4	
	Term	23461	23198	-	0
	Intr	24300	24198	-	0
10	Intr	24625	24565	-	0
	Init	24954	24707	-	0
	>2618599	/8490			
15	len =	1570	nex =	5	
	Term	27344	27150	-	0
	Intr	27528	27425	-	0
	Intr	27900	27615	-	0
20	Intr	28171	27989	-	0
	Init	28711	28527	-	0
	>2618599	/96			
25	len =	1631	nex =	5	
	Term	27344	27085	-	0
	Intr	27528	27425	-	0
	Intr	27900	27615	-	0
30	Intr	28171	27989	-	0
	Init	28715	28527	-	0
	>2618599	/96124			
35	len =	490	nex =	1	
	Sngl	61246	60764	-	0
	>2618599	/13096			
40	len =	1229	nex =	3	
	Init	71559	71780	+	0
	Intr	72273	72395	+	0
45	Term	72494	72787	+	0
	>2618599	/93205			
50	len =	977	nex =	3	
	Init	71649	71780	+	0
	Intr	72273	72395	+	0
	Term	72494	72625	+	0
55	>2618599	/34629			
	len =	1114	nex =	3	
	Init	71649	71780	+	0
60	Intr	72273	72395	+	0